

Semiparametric Approach for Regression with Covariate Subject to Limit of Detection

Shengchun Kong and Bin Nan

December 9, 2014

Abstract

We consider generalized linear regression analysis with left-censored covariate due to the lower limit of detection. Complete case analysis by eliminating observations with values below limit of detection yields valid estimates for regression coefficients, but loses efficiency; substitution methods are biased; maximum likelihood method relies on parametric models for the unobservable tail probability distribution of such covariate, thus may suffer from model misspecification. To obtain robust and more efficient results, we propose a semiparametric likelihood-based approach for the estimation of regression parameters using an accelerated failure time model for the covariate subject to limit of detection. A two-stage estimation procedure is considered, where the conditional distribution of the covariate with limit of detection given other variables is estimated prior to maximizing the likelihood function. The proposed method outperforms the complete case analysis and the substitution methods as well in simulation studies. Technical conditions for desirable asymptotic properties are provided.

Key words: Accelerate failure time model; Censored covariate; Empirical process; Generalized linear models; Pseudo-likelihood estimation.

1 Introduction

Detection limit is a threshold below which measured values are not considered significantly different from background noise (Helsel, 2005). Hence, values measured below this threshold are unreliable. In environmental epidemiology, particularly exposure analysis, when exposure levels are low, measurement of chemicals has a large percentage falling below the limit of detection due to inadequate instrument sensitivity. This is a common issue in the National Health and Nutrition Examination Survey (Crainiceanu et al., 2008), where many exposure variables have large proportions of measurements below their limits of detection. For example, 27.8% of the urine arsenobetaine measures are below its limit of detection at $0.4 \mu\text{g/l}$ (Caldwell et al., 2009). In the Diabetes Prevention Program, 66 of the 301 eligible participants had their testosterone levels

below the detection limit of 8.0 ng/dl (Kim et al., 2012). In an analysis for the Michigan Bone Health and Metabolism Study, up to 66% of the 50 study participants had anti-Mullerian hormone below the limit of detection at 0.05ng/ml (Sowers et al., 2008). For illustrative purpose in this article, we consider an analysis for the National Health and Nutrition Examination Survey, which examines the relationship between arsenic exposure and the prevalence of type 2 diabetes (Navas-Acien et al., 2008).

A variable with limit of detection can be either a response variable or a covariate in regression analysis. We focus on the latter in this article. Although many ad hoc methods have been implemented in practice, more appropriate statistical methods for regression models with a covariate subject to limit of detection are yet to be thoroughly studied (Schisterman and Little, 2010). The complete case analysis, simply eliminating observations with values below limit of detection, yields consistent estimates of the regression coefficients (Nie et al., 2010; Little and Rubin, 2002), but loses efficiency. Substitution methods are frequently used in epidemiology studies, where the values of covariate Z below limit of detection, denoted by L , are substituted by L , or $L/\sqrt{2}$, or zero, see for example, Schisterman et al. (2006), Hornung and Reed (1990), Moulton et al. (2002), Koru-Sengul et al. (2011), Kroger et al. (2009), Boomsma et al. (2009), Bloom et al. (2008), and Gollenberg et al. (2010) among many others. These methods are easily implementable, but found to be inappropriate and can yield large biases (Helsel, 2006; Nie et al., 2010). Richardson and Ciampi (2003) proposed to replace the values below limit of detection L with $E(Z|Z < L)$, which is obtained from an assumed known distribution of Z . There are two issues with this approach, however, (i) the distributional assumption is not verifiable; (ii) even if $E(Z|Z < L)$ is correctly specified, the method only leads to consistent estimates in linear regression when the covariate subject to limit of detection Z is independent of all the other covariates.

Another widely used method is the maximum likelihood estimation based on a parametric distributional assumption to the unobservable tail probability of the covariate that is subject to limit of detection. For examples, Nie et al. (2010) and Arunajadai and Rauh (2012) considered the linear regression based on a normal and a generalized gamma distribution for the covariate subject to limit of detection, respectively; Cole et al. (2009), Wu et al. (2012) and Albert et al. (2010) considered parametric maximum likelihood approach under the generalized linear regression; D'Angelo and Weissfeld (2008) applied this approach to the Cox regression. In practice, however, the underlying covariate distribution is unknown. The test of the parametric assumption to the unobservable tail probability of the covariate that is subject to limit of detection is usually unavailable because there is no information/observation below the detection limit. Both Lynn (2001) and Nie et al. (2010) noted that a parametric assumption can yield large bias if misspecified and argued that such an approach should not be attempted. Nie et al. (2010) recommended the complete case analysis despite the fact that simply dropping data below the limit of detection can lose a significant amount of information.

To obtain more efficient and yet robust results, we propose a semiparamet-

ric likelihood-based approach to fit generalized linear models with covariate subject to limit of detection. The tail distribution of the covariate beyond its limit of detection is estimated from a semiparametric accelerated failure model, conditional on all the fully observed covariates. Model checking can be done using martingale residuals for semiparametric accelerated failure time models. The proposed method is shown to be consistent and asymptotically normal, and outperforms existing methods in simulations. The proof of the asymptotic properties relies heavily on empirical process theory, which is provided in the online Supplementary Material.

2 A semiparametric approach

For a single observation, denote the response variable by Y , the covariate subject to limit of detection by Z , and the fully observed covariates by $X = (X_1, \dots, X_p)'$, where p is the number of fully observed covariates. For simplicity, we only consider one covariate that is subject to limit of detection. Consider a generalized linear model with

$$E(Y) = \mu = g^{-1}(D'\theta), \quad (1)$$

where g is the link function, $D'\theta$ is the linear predictor with $D = (1, X', Z)'$ and $\theta = (\beta', \gamma)'$, here β is a $(p+1)$ -dimensional vector and γ is a scalar. The variance of Y , typically a function of the mean, is denoted by

$$\text{var}(Y) = W(\mu) = W\{g^{-1}(D'\theta)\}.$$

We consider the exponential dispersion family in the natural form (Agresti, 2002; McCullagh and Nelder, 1989) given (Z, X)

$$f_{\varpi, \phi}(Y|Z, X) = \exp \left\{ \frac{Y\varpi - b(\varpi)}{a(\phi)} + c(Y, \phi) \right\}, \quad (2)$$

where ϕ is the dispersion parameter and ϖ is the natural parameter. We have $\mu = E(Y) = \dot{b}(\varpi)$, and $\text{var}(Y) = \ddot{b}(\varpi)a(\phi)$, where \dot{b} is the first derivative of b and \ddot{b} is the second derivative of b .

The actual value of Z is not observable when $Z < L$, where the constant L denotes the limit of detection, which is an example of left-censoring. In practice Z is a concentration measure of certain substance and thus non-negative. Consider a monotone decreasing transformation h that yields $Z = h(T)$, for example, $h(T) = \exp(-T)$. Denote $D(T) = (1, X', h(T))'$. If $T \leq C = h^{-1}(L)$, then T is observed; otherwise T is right-censored by C . We denote the observed value by $V = \min(T, C)$ and the censoring indicator by $\Delta = I(T \leq C)$.

The proposed methodology works for a broad family of link functions defined by the regularity conditions given in the Appendix. For notational simplicity, we present the main material using canonical link function g , where $g = (\dot{b})^{-1}$. Then, when T is observed, model (2) becomes

$$f_{\theta, \phi}(Y|T, X) = \exp \left\{ \frac{YD'(T)\theta - b(D'(T)\theta)}{a(\phi)} + c(Y, \phi) \right\}. \quad (3)$$

Denote the conditional cumulative distribution function of T given X by $F_1(t|X)$ with density $f_1(t|X)$. The likelihood function for the observed data (V, Δ, Y, X) can be factorized into

$$f(V, \Delta, Y, X) = f_2(V, \Delta|Y, X)f_3(Y|X)f_4(X),$$

where f denotes the joint density of (V, Δ, Y, X) , f_2 denotes conditional density of (V, Δ) given (Y, X) , f_3 denotes conditional density of Y given X , and f_4 denotes marginal density of X . Going through conditional arguments using the Bayes' rule and dropping $f_4(X)$, we obtain the likelihood function

$$L(V, \Delta, Y, X) = \{f_{\theta, \phi}(Y|T, X)f_1(T|X)\}^\Delta \left\{ \int_C^\infty f_{\theta, \phi}(Y|t, X)dF_1(t|X) \right\}^{1-\Delta}, \quad (4)$$

where only $f_{\theta, \phi}$ contains the parameter of interest θ , whereas f_1 is a nuisance parameter in addition to ϕ .

There are two parts in (4): (i) $\{f_{\theta, \phi}(Y|T, X)f_1(T|X)\}^\Delta$ for fully observed subject, and (ii) $\left\{ \int_C^\infty f_{\theta, \phi}(Y|t, X)dF_1(t|X) \right\}^{1-\Delta}$ for subject with covariate below limit of detection. Complete case analysis is only based on the first part and, although it yields a consistent estimate of θ , it clearly loses efficiency. We see from the second part of (4) that the efficiency gain comparing to the complete case analysis depends on how well we can recover the right tail of the conditional distribution $F_1(t|X)$ beyond C . Parametric models for $F_1(t|X)$ are often considered in the literature, see Nie et al. (2010), but it may suffer from model misspecification. The nonparametric method degenerates to the complete case analysis because there is no actual observation beyond censoring time C . We consider a semiparametric approach that allows reliable extrapolation beyond C and is robust against any parametric assumption.

Among all the commonly used semiparametric models for right-censored data, only the accelerated failure time model allows extrapolation beyond C , and model checking can be done by visualizing the cumulative sums of the martingale-based residuals (Lin et al., 1993, 1996; Peng and Fine, 2006). We hence propose a semiparametric accelerated failure time model for the transformed covariate subject to limit of detection given by

$$T = X'\alpha + \varsigma, \quad (5)$$

where ς follows some unknown distribution, denoted by η , and is independent of X . We only consider a fixed h for T in this article. More flexible transformation, for example, the Box-Cox transformation (Box and Cox, 1964; Foster et al., 2001; Cai et al., 2005), is worth further investigation. Note that X appears in both models (1) and (5), but it may refer to different forms of covariates in these models. For example, X_1 is a covariate in (1) whereas X_1^2 is a covariate in (5). We use the same X to denote all fully observed covariates for notational

simplicity. The log-likelihood function then becomes

$$\begin{aligned} \log L = & \Delta \log f_{\theta, \phi}(Y|T, X) + \Delta \log \dot{\eta}(T - X' \alpha) \\ & + (1 - \Delta) \log \left\{ \int_{C - X' \alpha}^{\tau} f_{\theta, \phi}(Y|t + X' \alpha, X) d\eta(t) \right\}, \end{aligned} \quad (6)$$

where τ is a truncation time at the residual scale defined in Condition 4 in the Appendix.

3 The pseudo-likelihood method

The log likelihood function (6) involves an unknown distribution function η and its derivative, hence a semiparametric maximum likelihood estimation, if it exists, can be complicated. We propose a tractable two-stage pseudo-likelihood approach in which the nuisance parameters (ϕ, α, η) are estimated in stage 1, and the parameter of interest θ is then estimated by maximizing the data version of (6) in stage 2 with nuisance parameters replaced by their estimators obtained in stage 1. Details are given below:

Stage 1. Nuisance parameter estimation. Dispersion parameter ϕ is estimated by the complete case analysis of the generalized linear model (2); the accelerated failure time model regression coefficient α is estimated by either the rank based methods, see Wei et al. (1990), Jin et al. (2003), Nan et al. (2009) or the sieve maximum likelihood method, see Ding and Nan (2011); and the accelerated failure time model error distribution η is estimated by the Kaplan-Meier estimator from the censored residuals.

The complete case analysis can be done by any standard statistical package for generalized linear models. The rank based estimates for the accelerated failure time model usually are obtained by using linear programming. The R package “rankreg” (Zhou, 2006), now archived by CRAN, can be implemented for small to moderate sample sizes because it solves a linear programming problem of size n^2 . This is the method we implemented in simulations and the arsenic exposure data example. An alternative approach is to modify the Newton algorithm for solving the discrete rank based estimating equation (Yu and Nan, 2006). Standard Newton-Raphson algorithm can be implemented to obtain the sieve maximum likelihood estimates (Ding and Nan, 2011) for the accelerated failure time model when the sample size is large.

Stage 2. Pseudo-likelihood estimation of θ . Replacing (ϕ, α, η) by their Stage 1 estimates $(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n})$ in the log likelihood function yields the following log pseudo-likelihood function for a random sample of n observations:

$$\begin{aligned} pl_n(\theta) = & \frac{1}{n} \sum_{i=1}^n \left\{ \Delta_i \log f_{\theta, \hat{\phi}_n}(Y_i|X_i, T_i) \right. \\ & \left. + (1 - \Delta_i) \log \int_{C - X_i' \hat{\alpha}_n}^{\tau} f_{\theta, \hat{\phi}_n}(Y_i|X_i, t + X_i' \hat{\alpha}_n) d\hat{\eta}_{n, \hat{\alpha}_n}(t) \right\}, \end{aligned} \quad (7)$$

where

$$f_{\theta, \hat{\phi}_n}(Y_i|T_i, X_i) = \exp \left[\frac{Y_i \{D'_i(T_i)\theta\} - b\{D'_i(T_i)\theta\}}{a(\hat{\phi}_n)} + c(Y_i, \hat{\phi}_n) \right].$$

Note that the term $\Delta \log \dot{\eta}(T)$ in (6) is dropped because it does not involve θ . We maximize (7) by setting its derivative to zero and then solving the equation to obtain the pseudo-likelihood estimator $\hat{\theta}_n$. This is implemented by using standard Newton-Raphson algorithm with the initial value obtained from the complete case analysis in Stage 1.

Since $\hat{\theta}_n$ is obtained by solving an estimating equation, its asymptotic properties can be obtained from Z-estimation theory. It can be shown that all the estimates obtained in Stage 1 have desirable statistical properties for Stage 2 estimation. In particular, $\hat{\phi}_n$ obtained from the complete case analysis is $n^{1/2}$ -consistent by Little and Rubin (2002); $\hat{\alpha}_n$ is $n^{1/2}$ -consistent by Nan et al. (2009) or Ding and Nan (2011); and $\hat{\eta}_{n, \hat{\alpha}_n}$ is also $n^{1/2}$ -consistent in a finite interval, and its proof is provided in the online Supplementary Material.

4 Asymptotic properties

Define a random map as follows

$$\Psi_{\theta, n}(\phi, \alpha, \eta) = \frac{1}{n} \sum_{i=1}^n \psi_{\theta}(Y_i, X_i, V_i, \Delta_i; \phi, \alpha, \eta), \quad (8)$$

where

$$\begin{aligned} \psi_{\theta}(Y, X, V, \Delta; \phi, \alpha, \eta) &= \Delta \{Y - \dot{b}(D'(T)\theta)\} D(T) + (1 - \Delta) \left\{ \int_{C-X'\alpha}^{\tau} f_{\theta, \phi}(Y|t + X'\alpha, X) d\eta(t) \right\}^{-1} \\ &\quad \int_{C-X'\alpha}^{\tau} f_{\theta, \phi}(Y|t + X'\alpha, X) \{Y - \dot{b}(D'(t + X'\alpha)\theta)\} D(t + X'\alpha) d\eta(t), \end{aligned}$$

which is the derivative of (6) with respect to θ . Then with (ϕ, α, η) replaced by $(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n})$ in (8), $\Psi_{\theta, n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n}) = 0$ becomes the pseudo-likelihood estimating equation for θ , and its solution $\hat{\theta}_n$ is called the pseudo-likelihood estimator.

A set of regularity conditions is introduced in the Appendix. Some conditions are commonly assumed for the accelerated failure time models, and other conditions are for the generalized linear models, which are easily verifiable for linear, logistic and Poisson regression models. We then have the following asymptotic results for $\hat{\theta}_n$.

Theorem 4.1. (*Consistency and asymptotic normality.*) Denote the true value of θ by θ_0 . Suppose all the regularity conditions given in the Appendix hold.

Then for the two-stage pseudo-likelihood estimator $\hat{\theta}_n$ satisfying $\Psi_{\hat{\theta}_n, n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n}) = 0$, we have: (i) $\hat{\theta}_n$ converges in outer probability to θ_0 , and (ii) $n^{1/2}(\hat{\theta}_n - \theta_0)$ converges weakly to a mean zero normal random variable with variance $A^{-1}BA^{-1}$, where A and B are provided in the online Supplementary Material.

Because the asymptotic variance of $\hat{\theta}_n$ has a very complicated expression that prohibits the direct calculation of its estimate from observed data, we recommend using the bootstrap variance estimator.

The proof of Theorem 4.1 is based on the general Z-estimation theory of Nan and Wellner (2013). Define a deterministic function

$$\Psi_{\theta}(\phi, \alpha, \eta) = E \left\{ \psi_{\theta}(Y, X, V, \Delta; \phi, \alpha, \eta) \right\}, \quad (9)$$

and denote the true values of (ϕ, α, η) by $(\phi_0, \alpha_0, \eta_0)$. We can show that $\Psi_{\theta, n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n})$ converges uniformly to $\Psi_{\theta}(\phi_0, \alpha_0, \eta_0)$ as $n \rightarrow \infty$. Then the consistency is achieved given that θ_0 is the unique solution of $\Psi_{\theta}(\phi_0, \alpha_0, \eta_0) = 0$. The asymptotic normality is derived by showing the asymptotic linear representation of $n^{1/2}(\hat{\theta}_n - \theta_0)$. The detailed proofs rely heavily on empirical process theory and can be found in the online Supplementary Material, where we only provide the analytic form of the asymptotic variance for the Gehan weighted estimate of α . The analytic forms of the asymptotic variance for other rank based estimates and the sieve maximum likelihood estimates can be obtained similarly.

5 Numerical results

5.1 Simulations

We conduct simulations to investigate the finite sample performance of the proposed method. Simulation data sets are generated from the generalized linear model

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma Z,$$

where $\beta_0 = -1$, $\beta_1 = 0.5$, $\beta_2 = -1$, $\gamma = 2$, and g is chosen to be the canonical link function for normal, bernoulli and poisson distributions, respectively. The normal error variance is chosen to be 1 for the linear regression model. The three covariates are: $X_1 \sim \text{Bernoulli}(0.5)$, X_2 is normal with mean 1 and standard deviation 1 truncated at ± 3 , and $Z = \exp(-T)$ is generated from the following linear model

$$T = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \varsigma,$$

where $\alpha_0 = 0.25$, $\alpha_1 = 0.25$, $\alpha_2 = -0.5$, $\varsigma \sim 0.5N(0, 1/8^2) + 0.5N(0.5, 1/10^2)$. The limit of detection L for covariate Z is chosen to yield 30% right censoring for T .

We simulate 1000 replications for each scenario and compare the proposed method with full data analysis, complete case analysis, and four different substitution methods. The full data analysis represents the case of no limit of

detection, which serves as a benchmark. We conduct simulations with two different sample sizes: 200 and 400. The four substitution methods for $Z < L$ are: (i) replacing Z by L , (ii) replacing Z by $L/\sqrt{2}$, (iii) replacing Z by zero, and (iv) replacing Z by $E(Z|Z < L)$. For the proposed two-stage method, we report the 90% and 95% coverage proportions for which the variances are obtained from 200 bootstrap samples. Empirical variances obtained from 1000 independent data sets are provided only for the full data analysis and the two valid methods for the limit of detection problem. The results are presented in Tables 1-3.

The results suggest that all the substitution methods yield biased estimates, including substituting Z by the true $E(Z|Z < L)$. The biases for the proposed two-stage method are minimal, which are comparable to both the full data analysis and the complete case analysis. Clearly, the proposed method is much more efficient than the complete case analysis, and the bootstrap method performs well in estimating the variance, which yields reasonable coverage rate of the confidence intervals for all considered sample sizes.

Different censoring rates varying from 10% to 60% were considered in additional simulations. Both the proposed two-stage method and the complete case analysis yield unbiased estimates in all simulation scenarios. The efficiency gain of the two-stage method comparing to complete case analysis increases as the percentage of censoring increases. All the substitution methods yield biased estimates and the biases increase as the censoring rate increases as well. The detailed simulation results are not provided here.

The consistency of the estimates from the two-stage method depends on correctly specifying the semiparametric accelerated failure time model (5). Although the assumption for model (5) is much less restrictive than any parametric model, model checking should be done before applying the proposed two-stage method. Though less severe than ad hoc substitution methods, misspecification of the accelerated failure time model also yields biased results, and the bias increases as the severity of misspecification of (5) grows. Again, the detailed simulation results are not provided here.

5.2 The National Health and Nutrition Examination Survey

We consider the National Health and Nutrition Examination Survey 2003-2004 as an illustrative example for the regression with a covariate subject to limit of detection. In particular, we focus on the effect of left-censored urine arsenobetaine on the prevalence of type 2 diabetes, see Navas-Acien et al. (2008)

National Health and Nutrition Examination Survey, conducted by the US National Center for Health Statistics, used a complex multistage sampling design to obtain a representative sample of the civilian noninstitutionalized individuals within the US population. The data set contains a subsample of 1542 study participants with arsenic measurements of the National Health and Nutrition Examination Survey 2003-2004. For each participant in this subsample, total urine arsenic and arsenic species including arsenobetaine were collected for

Table 1: Simulation results for linear regression.

	Sample size		$\beta_0 = -1$	$\beta_1 = 0.5$	$\beta_2 = -1$	$\gamma = 2$
Full data	200	bias	-0.030	0.011	-0.006	0.024
		var	0.414	0.051	0.036	0.374
Two-stage		bias	-0.029	0.010	-0.005	0.022
		var	0.438	0.053	0.038	0.395
Complete case		bootstrap var	0.465	0.058	0.043	0.421
		90% CR (%)	89.0	90.7	91.2	89.0
		95% CR (%)	94.9	94.8	95.6	95.3
		Complete case	bias	-0.018	0.004	0.000
var			0.531	0.082	0.066	0.507
L		bias	0.399	-0.220	0.228	-0.491
$L/\sqrt{2}$	bias	0.701	-0.136	0.147	-0.620	
Zero	bias	1.837	-0.417	0.427	-1.684	
$E(Z Z < L)$	bias	0.415	-0.133	0.143	-0.418	
Full data	400	bias	-0.019	0.007	-0.003	0.014
		var	0.212	0.028	0.019	0.192
Two-stage		bias	-0.019	0.008	-0.003	0.015
		var	0.225	0.029	0.020	0.204
Complete case		bootstrap var	0.226	0.028	0.021	0.205
		90% CR (%)	89.4	89.2	90.4	89.9
		95% CR (%)	95.0	93.8	95.0	95.0
		Complete case	bias	-0.019	-0.001	0.004
var			0.273	0.043	0.033	0.255
L		bias	0.404	-0.221	0.230	-0.495
$L/\sqrt{2}$	bias	0.724	-0.144	0.155	-0.642	
Zero	bias	1.850	-0.426	0.436	-1.700	
$E(Z Z < L)$	bias	0.433	-0.138	0.148	-0.434	

Table 2: Simulation results for logistic regression.

	Sample size		$\beta_0 = -1$	$\beta_1 = 0.5$	$\beta_2 = -1$	$\gamma = 2$
Full data	200	bias	-0.030	0.013	-0.033	0.060
		var	2.157	0.268	0.216	2.033
Two-stage		bias	-0.041	0.016	-0.037	0.071
		var	2.313	0.278	0.230	2.191
		bootstrap var	2.424	0.299	0.235	2.260
		90% CR (%)	91.4	91.7	90.7	91.0
		95% CR (%)	96.2	96.3	95.9	96.2
Complete case		bias	-0.076	0.021	-0.045	0.106
		var	2.822	0.413	0.381	2.842
L		bias	0.309	-0.185	0.171	-0.368
$L/\sqrt{2}$		bias	0.690	-0.122	0.110	-0.570
Zero		bias	1.880	-0.453	0.441	-1.716
$E(Z Z < L)$		bias	0.350	-0.100	0.087	-0.313
Full data	400	bias	-0.033	0.007	-0.016	0.041
		var	0.930	0.123	0.096	0.881
Two-stage		bias	-0.043	0.011	-0.020	0.052
		var	1.013	0.129	0.104	0.964
		bootstrap var	1.101	0.138	0.107	1.022
		90% CR (%)	90.8	91.2	90.5	90.6
		95% CR (%)	95.8	96.3	95.8	95.2
Complete case		bias	-0.037	0.005	-0.018	0.048
		var	1.169	0.190	0.159	1.160
L		bias	0.319	-0.193	0.190	-0.398
$L/\sqrt{2}$		bias	0.651	-0.119	0.117	-0.553
Zero		bias	1.841	-0.442	0.440	-1.691
$E(Z Z < L)$		bias	0.332	-0.103	0.101	-0.317

Table 3: Simulation results for Poisson regression.

	Sample size		$\beta_0 = -1$	$\beta_1 = 0.5$	$\beta_2 = -1$	$\gamma = 2$
Full data	200	bias	0.022	-0.009	0.008	-0.026
		var	0.225	0.024	0.018	0.198
Two-stage		bias	0.034	-0.011	0.011	-0.037
		var	0.250	0.026	0.020	0.221
		bootstrap var	0.249	0.027	0.020	0.218
		90% CR (%)	90.9	89.9	90.0	90.6
		95% CR (%)	94.5	94.8	94.7	94.8
Complete case		bias	0.025	-0.011	0.010	-0.031
		var	0.351	0.053	0.041	0.325
L		bias	0.589	-0.288	0.286	-0.660
$L/\sqrt{2}$		bias	0.885	-0.200	0.210	-0.801
Zero		bias	1.867	-0.380	0.396	-1.691
$E(Z Z < L)$		bias	0.637	-0.213	0.217	-0.628
Full data	400	bias	0.018	-0.003	0.005	-0.020
		var	0.105	0.012	0.008	0.092
Two-stage		bias	0.019	-0.003	0.005	-0.021
		var	0.119	0.013	0.009	0.104
		bootstrap var	0.121	0.013	0.010	0.105
		90% CR (%)	90.1	90.7	90.5	90.7
		95% CR (%)	95.2	95.3	94.8	95.0
Complete case		bias	0.016	-0.004	0.007	-0.022
		var	0.175	0.027	0.022	0.163
L		bias	0.578	-0.281	0.283	-0.649
$L/\sqrt{2}$		bias	0.886	-0.196	0.208	-0.800
Zero		bias	1.870	-0.373	0.391	-1.689
$E(Z Z < L)$		bias	0.633	-0.208	0.215	-0.623

arsenic analysis. The limit of detection for total urine arsenic and urine arsenobetaine were 0.6 and 0.4 $\mu\text{g/l}$, respectively. The percentage of study participants with levels below the limit of detection were 1.3% for total urine arsenic and 27.8% for urine arsenobetaine (Caldwell et al., 2009). Navas-Acien et al. (2008) found that total urine arsenic was associated with increased prevalence of type 2 diabetes, hence it was adjusted in the analysis, and 19 participants with total urine arsenic below limit of detection were dropped from the study. The urine creatinine level that was used to account for urine dilution in spot urine samples was fully observed and was adjusted in the analysis as well. We further excluded 23 participants with missing values in other variables of interest.

For illustration purpose, we focus on the male participants, where 24.1% of 730 subjects had urine arsenobetaine below limit of detection. Age, race/ethnicity, body mass index, the logarithm of total urine arsenic and the logarithm of urine creatinine level are used as covariates to fit the accelerated failure time model for $-\log(\text{arsenobetaine})$. All of them are significant with p-values < 0.0001 except one dummy variable for race. Dotted lines in Figure 1 are the plots of 50 realizations from the distributions of the score processes. The observed score processes are presented with solid lines which randomly fluctuate around zero. From Figure 1 we see that the considered accelerated failure time model for urine arsenobetaine fits the data reasonably well, with respective goodness-of-fit empirical p-values of 0.686, 0.104, 0.706, 0.782, 0.68, 0.794, 0.646 and 0.834 for age, log total urine arsenic, log urine creatinine, body mass index and race that contains five ethnic groups based on 500 simulated martingale residual score processes. The response of interest is the status of type 2 diabetes. Thus, a logistic regression is considered with age, body mass index, log total urine arsenic, log urine creatinine and log urine arsenobetaine as the covariates of interest, whereas race is not significantly associated. Table 4 shows the regression coefficient estimates, where we see that the proposed two-stage method yields similar point estimates with smaller variances/p-values comparing to the complete case analysis, indicating the efficiency gain of the proposed method. Substitution methods with values below limit of detection replaced by L or $L/\sqrt{2}$ yield quite different point estimates comparing to the complete case analysis and the two-stage method, thus put the significant results of $\log(\text{arsenobetaine})$ in question. The effect of $\log(\text{arsenobetaine})$ is clearly biased when the urine arsenobetaine values below limit of detection are replaced by zero.

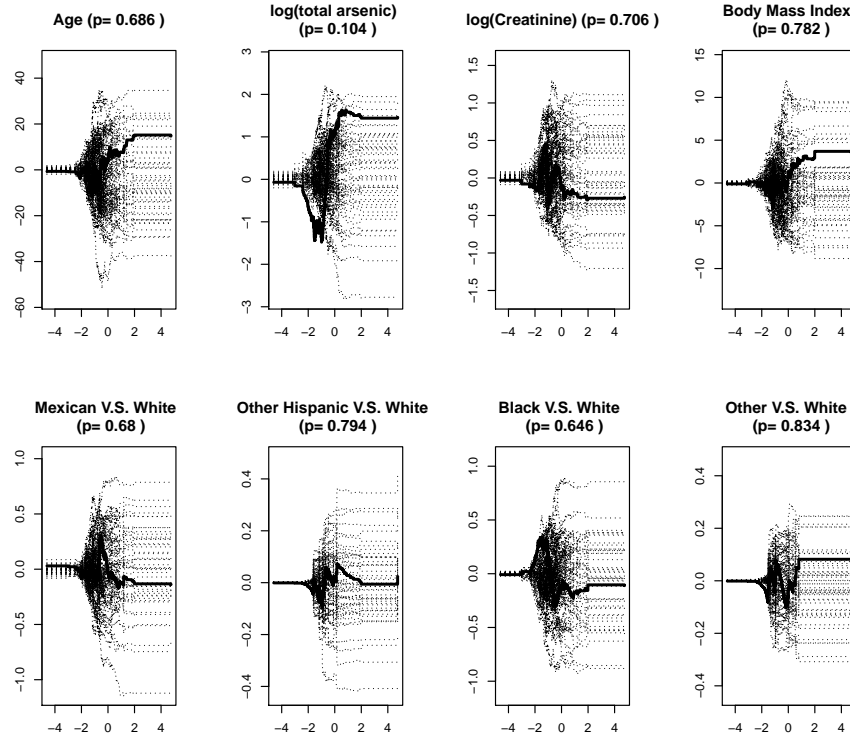
6 Discussion

The estimates from the proposed method are consistent and asymptotically normal under much less restrictive assumptions than parametric approaches. Another advantage of the two-stage method over the parametric method is that the model checking tools are available. All the substitution methods could yield large bias including replacing the values below the limit of detection by the true $E(Z|Z < L)$. The consistency and efficiency gain of the proposed two-

Table 4: Regression analysis results for the prevalence of type 2 diabetes with co-variate urinary arsenobetaine subject to limit of detection: the National Health and Nutrition Examination Survey 2003-2004.

		Age	log(Arsenic)	log(Creatinine)	BMI	log(Arsenobetaine)
Two-stage	estimate	0.05	0.50	-0.76	0.10	-0.22
	bootstrap sd	0.007	0.25	0.27	0.02	0.12
	p-value	< 0.0010	0.04	0.004	< 0.0001	0.07
Complete case	estimate	0.06	0.41	-0.81	0.13	-0.20
	sd	0.01	0.32	0.33	0.03	0.20
	p-value	< 0.0010	0.20	0.01	< 0.0001	0.31
L	estimate	0.05	0.57	-0.81	0.10	-0.32
	sd	0.008	0.25	0.26	0.02	0.15
	p-value	< 0.0001	0.02	0.002	< 0.0001	0.03
$L/\sqrt{2}$	estimate	0.05	0.56	-0.80	0.10	-0.30
	sd	0.008	0.24	0.26	0.02	0.14
	p-value	< 0.0001	0.02	0.002	< 0.0001	0.02
Zero	estimate	0.05	0.25	-0.61	0.10	-0.03
	sd	0.008	0.16	0.24	0.02	0.02
	p-value	< 0.0001	0.12	0.009	< 0.0001	0.11

Figure 1: Goodness of fit for the AFT model



stage method rely on the correctly specified accelerated failure time model. We suggest to fit the accelerated failure time model for the covariate subject to limit of detection before applying the two-stage method. If the data fits the model reasonably well and there are at least some fully observed covariates significantly associated with the covariate subject to limit of detection, then the proposed method is recommended. Otherwise, we suggest to use the complete case analysis.

The amount of efficiency gain of the proposed two-stage method depends on how far we can estimate $F(t|X)$ reasonably well beyond the limit of detection. We assume some finite value τ for residuals in this article. In practice, the upper limit of the integral in the pseudo-likelihood function can go as far as the largest observed residual in the fitted accelerated failure time model, which is $\max_i(T_i - X_i'\hat{\alpha}_n)$; and theoretically, this upper limit is ∞ when the support of $X'\alpha_0$ is unbounded. In the latter case, it can be shown that \hat{F}_n converges to F on the entire real line with a polynomial rate at $n^{-1/8}$, see Lai and Ying (1991) and Ding and Nan (2014), and we may still obtain consistent estimates for the parameters of interest. The asymptotic normality, however, will be difficult to show.

We only consider the case with one covariate subject to limit of detection in this article for simplicity. Regression with multiple covariates subject to limits of detection may occur in practice. Parametric models have been considered for such problems (May et al., 2011; D'Angelo and Weissfeld, 2008). To achieve robust results, the proposed semiparametric approach can be generalized to tackle the problem with multiple covariates subject to limits of detection. The critical step is to provide a valid nonparametric estimate for the multivariate survival function, for which available methods include Dabrowska (1988), Prentice and Cai (1992), van der Laan (1996), and Prentice and Moodie (2004). The constant limit of detection assumption considered in this article, though commonly seen in practice, is also for notational simplicity, and can be relaxed to cases with random limit of detection.

Limit of detection issue is a special missing data problem. Multiple imputation (Little and Rubin, 2002) may be considered as an alternative method if the tail distribution of the covariate subject to limit of detection conditional on all other variables, including the response variable, can be estimated reasonably well. This research is ongoing and will be presented elsewhere.

When the original lab measurements are available, which may or may not be below limit of detection, Murphy et al. (2010) and Buck Louis et al. (2012) directly used the machine-read lab values in their analysis to avoid the potential biases caused by substitution. More appropriate analysis should be treating these lab data as error prone measurements (Guo and Little, 2010), hence methods dealing with measurement errors would apply.

The proposed two-stage method is ready to be generalized to other regression models that have a likelihood function to work with, for example, the Cox regression model and mixed effects model. Extra care will be needed due to special features of these models. For example, handling the nonparametric

baseline function in the Cox model in the context of limit of detection can be delicate. These are interesting topics worth further investigation.

Supplementary material

The online Supplementary Material contains the detailed proof of Theorem 1.

Acknowledgement

We thank Professor Sung Kyun Park for the National Health and Nutrition Examination Survey arsenic exposure data. We are grateful to the editor, associate editor and a referee for helpful and constructive comments. The research was partially supported by several grants from the U.S. National Science Foundation and the National Institute of Health.

Appendix: regularity conditions

Denote the sample space of response variable Y by \mathcal{Y} , the sample space of covariate X by \mathcal{X} , the parameter space of θ by Θ , the parameter space of α by \mathcal{A} , and the parameter space of η by \mathcal{H} . In addition to the assumptions of bounded support for (X, Z) and compact parameter spaces Θ and \mathcal{A} , we provide a set of regularity conditions for Theorem 4.1 in the following.

Condition 1. $\Psi_\theta(\phi_0, \alpha_0, \eta_{0, \alpha_0})$ has a unique root θ_0 .

Condition 2. For any constant $U < \infty$, $\sup_{t \in [C, U]} |h(t)| \leq E_0 < \infty$, $\sup_{t \in [C, U]} |\dot{h}(t)| \leq E_1 < \infty$, and $\sup_{t \in [C, U]} |\ddot{h}(t)| \leq E_2 < \infty$, where \dot{h} and \ddot{h} are the first and second derivatives of h respectively, and E_0 , E_1 and E_2 are constants.

Condition 3. Error ς has bounded density $f = \dot{\eta}_{0, \alpha_0}$ with bounded derivative \dot{f} , in other words, $f \leq E_3 < \infty$, $|\dot{f}| \leq E_4 < \infty$ for constants E_3 and E_4 , and

$$\int_{-\infty}^{\infty} (\dot{f}(t)/f(t))^2 f(t) dt < \infty.$$

Condition 4. There is a constant $\tau < \infty$ such that $\text{pr}(V - X'\alpha \geq \tau | X = x) > \xi > 0$ for all $x \in \mathcal{X}$ and $\alpha \in \mathcal{A}$.

Condition 5. $a(\phi)$ is a monotone function satisfying $|1/a(\phi)| \leq l < \infty$ for a constant l with bounded derivatives $\dot{a}(\cdot)$ and $\ddot{a}(\cdot)$.

Condition 6. $\dot{b}(\cdot)$ is a bounded monotone function.

Condition 7. $\ddot{b}(\cdot)$ is a bounded Lipschitz function.

Condition 8. *There exist constants C_i , $i = 1, \dots, 5$, such that for any constant $U < \infty$,*

$$\begin{aligned} \sup_{y \in \mathcal{Y}, \theta \in \Theta, |1/a(\phi)| \leq l, x \in \mathcal{X}, t \in [C, U]} \left| f_{\theta, \phi}(y|t, x) \{y - \dot{b}(D'(t)\theta)\} \right| &\leq C_1 < \infty, \\ \sup_{y \in \mathcal{Y}, \theta \in \Theta, |1/a(\phi)| \leq l, x \in \mathcal{X}, t \in [C, U]} \left| \frac{\partial f_{\theta, \phi}(y|t, x)}{\partial \phi} \{y - \dot{b}(D'(t)\theta)\} \right| &\leq C_2 < \infty, \\ \sup_{y \in \mathcal{Y}, \theta \in \Theta, |1/a(\phi)| \leq l, x \in \mathcal{X}, t \in [C, U]} \left| \frac{\partial \left[f_{\theta, \phi}(y|t, x) \{y - \dot{b}(D'(t)\theta)\} \right]}{\partial t} \right| &\leq C_3 < \infty, \\ \sup_{y \in \mathcal{Y}, \theta \in \Theta, |1/a(\phi)| \leq l, x \in \mathcal{X}, t \in [C, U]} \left| \frac{\partial f_{\theta, \phi}(y|t, x)}{\partial \phi} \right| &\leq C_4 < \infty, \\ \sup_{y \in \mathcal{Y}, \theta \in \Theta, |1/a(\phi)| \leq l, x \in \mathcal{X}, t \in [C, U]} \left| \frac{\partial \left[f_{\theta, \phi}(y|t, x) \{y - \dot{b}(D'(t)\theta)\} \right]}{\partial \theta} \right| &\leq C_5 < \infty. \end{aligned}$$

Condition 9. *There exist constants $\delta_1 > 0$ and $\delta_2 > 0$, such that $\int_{C-X'\alpha}^{\tau} f_{\theta, \phi}(Y|t + X'\alpha, X) d\eta(t) \geq \delta_1$ with probability 1 for any $\theta \in \Theta$ and $|\phi - \phi_0| + |\alpha - \alpha_0| + \|\eta - \eta_0\| < \delta_2$.*

REMARK: Condition 1 is for the consistency, which may be unnecessarily strong for the proposed two-stage method. Direct calculation yields

$$\begin{aligned} \dot{\Psi}_{\theta_0} &= \left. \frac{\partial \Psi_{\theta}(\phi_0, \alpha_0, \eta_0)}{\partial \theta} \right|_{\theta=\theta_0} \\ &= E \left\{ -\Delta \ddot{b}\{D'(T)\theta_0\} D(T)^{\otimes 2} - (1 - \Delta) \left(\int_{C-X'\alpha_0}^{\tau} f_{\theta_0, \phi_0}(Y|t + X'\alpha_0, X) d\eta_0(t) \right)^{-2} \right. \\ &\quad \left. \left(\int_{C-X'\alpha_0}^{\tau} f_{\theta_0, \phi_0}(Y|t + X'\alpha_0, X) [Y - \dot{b}\{D'(t + X'\alpha_0)\theta_0\}] D(t + X'\alpha_0) d\eta_0(t) \right)^{\otimes 2} \right\}, \end{aligned}$$

which is negative definite. Thus $\dot{\Psi}_{\theta}$, a continuous matrix with θ , is also negative definite in a neighborhood of θ_0 , which guarantees that θ_0 is the unique solution of $\Psi_{\theta}(\phi_0, \alpha_0, \eta_0) = 0$ in a neighborhood of θ_0 . The initial value we use in the Newton-Raphson algorithm for solving $\Psi_{\theta, n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_n, \hat{\alpha}_n) = 0$ is obtained from the complete case analysis, which is consistent, thus the solution of the proposed two-stage method should also be consistent.

Condition 2 holds for many commonly used transformations, for example, $h(t) = \exp(-t)$ and polynomial functions. Condition 3 and 4 are usual assumptions for accelerated failure time models (Tsiatis, 1990; Nan et al., 2009). Conditions 5-8 automatically hold for common generalized linear models, for example, linear, logistic or poisson regression.

Condition 9 is mainly for technical convenience. One way to obtain Condition 9 might be to truncate response variable Y such that $|Y| \leq M < \infty$ for a large constant M and to further truncate the residual in the accelerated failure

time model with some constant $\tau' < \tau$. In our simulations, however, we do not implement such truncations but still obtain satisfactory results.

References

- AGRESTI, A. (2002). Categorical data analysis. *John Wiley*.
- ALBERT, P.S., HAREL, O., PERKINS, N. & BROWNE, R. (2010). Use of Multiple Assays Subject to Detection Limits With Regression Modeling in Assessing the Relationship Between Exposure and Outcome. *Epidemiology* **21**, S35-43.
- ARUNAJADAI, S.G. & RAUH, V.A. (2012). Handling covariates subject to limits of detection in regression . *Environmental and Ecological Statistics* **19**, 369-391.
- BLOOM, M.S., VORN SAAL, F.S., KIM, D., TAYLOR, J.A., LAMB, J.D. & FUJIMOTO, V.Y. (2008). Environmental exposure to PBDEs and thyroid function among New York anglers. *Environmental Toxicology and Pharmacology* **32**, 319-323.
- BOOMSMA C.M., KAVELAARS, A., EIJKEMANS, MJC., AMAROUCHI, K., TEKLENBURG, G., GUTKNECHT, D., FAUSER, BJCM., HEIJNEN, C.J. & MACKLON, N.S. (2009). Cytokine profiling in endometrial secretions: a non-invasive window on endometrial receptivity. *Reproductive BioMedicine* **18**, 85-94.
- BOX, G. E. P. & COX, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B.* **26**, 211-252.
- BUCK LOUIS, G.M., CHEN, Z., PETERSON, C.M., HEDIGER, M.L., CROUGHAN, M.S., SUNDARAM, R., STANFORD, J.B., VARNER, M.W., FUJIMOTO, V.Y., GIUDICE, L.C., TRUMBLE, A., PARSONS, P.J. & KANNAN, K. (2012). Persistent Lipophilic Environmental Chemicals and Endometriosis: The ENDO Study. *Environ Health Perspect.* **120**, 811-816.
- CAI, T., TIAN, L. & WEI, L. J. (2005). Semiparametric Box-Cox power transformation models for censored survival observations. *Biometrika* **92**, 619-632.
- CALDWELL, K.L., JONES, R.L., VERDON, C.P., JARRETT, J.M., CAUDILL, S.P. & OSTERLOH, J.D. (2009). Levels of urinary total and speciated arsenic in the US population: National Health and Nutrition Examination Survey 2003C2004. *Journal of Exposure Science and Environmental Epidemiology* **19**, 59-68.
- CHEN, H., QUANDT, S.A., GRZYWACZ, J.G. & ARCURY, T.A. (2011). A Distribution-Based Multiple Imputation Method for Handling Bivariate Pesticide Data with Values below the Limit of Detection. *Environmental Health Perspectives* **119**, 351-356.

- CRAINICEANU, C.M., GUALLAR, E., NAVAS-ACIEN, A. & TELLEZ-PLAZA, M. (2008). Cadmium exposure and hypertension in the 1999-2004 National Health and Nutrition Examination Survey (Nutrition Examination Survey). *Environmental Health Perspectives* **116**, 51-56.
- COLE, S.R., CHU, H., NIE, L. & SCHISTERMAN, E.F. (2009). Estimating the odds ratio when exposure has a limit of detection. *International Journal of Epidemiology* **38**, 1674-1680.
- DABROWSKA, D.M. (1988). Kaplan-Meier estimate on the plane. *Annals of Statistics* **16**, 1475-1489.
- D'ANGELO, G. & WEISSFELD, L. (2008). An index approach for the Cox model with left censored covariates. *Statistics in medicine* **72**, 4502-4514.
- DING, Y. & NAN, B. (2011). A sieve M-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *Annals of Statistics* **39**, 3032-3061.
- DING, Y. & NAN, B. (2014). You have full text access to this content Estimating Mean Survival Time: When is it Possible? *Scandinavian Journal of Statistics*, in press.
- FOSTER, A. M., TIAN, L. & WEI, L. J. (2001). Estimation for the Box-Cox transformation model without assuming parametric error distribution. *Journal of the American Statistical Association* **96**, 1097-1101.
- GILLESPIE, B.W., CHEN, Q., REICHERT, H., FRANZBLAU, A., HEDGEMAN, E., LEPKOWSKI, J., ADRIAENS, P., DEMOND, A., LUKSEMBURG, W. & GARABRANDT, D.H. (2010). Estimating Population Distributions When Some Data Are Below a Limit of Detection by Using a Reverse Kaplan-Meier Estimator. *Epidemiology* **21**, S64CS70.
- GOLLENBERG, A.L., HEDIGER, M.L., LEE, P.A., HIMES, J.H. & BUCK LOUIS, G.M. (2010). Association between Lead and Cadmium and Reproductive Hormones in Peripubertal U.S. Girls. *Environ Health Perspect* **118**, 1782-1787.
- GUO, Y., HAREL, O. & LITTLE, R.J. (2010). How well quantified is the limit of quantification? *Epidemiology*. **21**(4) S10-S16.
- HELSEL, D. R. (2005). Nondetects and Data Analysis: Statistics for Censored Environmental Data. New York: Wiley .
- HELSEL, D. R. (2006). Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere* **56**, 2434-2439.
- HORNUNG, R.W. & REED, L.D. (1990). Estimation of average concentration in the presence of nondetectable values. *Applied Occupational and Environmental Hygiene* **4**, 46-51.

- JAIN, R.B., CAUDILL, S.P., WANG, R.Y. & MONSELL, E. (2008). Evaluation of Maximum Likelihood Procedures To Estimate Left Censored Observations. *Anal. Chem.* **80**, 1124-1132.
- JIN, Z., LIN, D. Y., WEI, L. J. & YING, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341-353.
- KIM, C., KONG, S., LAUGHLIN, G. A., GOLDEN, S. H., MATHER, K. J., NAN, B., RANDOLPH, J. R., EDELSTEIN, S. L., LABRIE, F., BUSCHUR E. & BARRETT-CONNOR, E. ^a (2012). Reductions in glucose among post-menopausal women who use and do not use estrogen therapy. *Menopause* **20**, 4.
- KROGER, E., VERRAULT, R., CARMICHAEL, P., LINDSAY, J., JULIEN, P., DEWAILLY, E., AYOTTE, P. & LAURIN, D. (2009). Omega-3 fatty acids and risk of dementia: the Canadian Study of Health and Aging1-4. *Am J Clin Nutr* **90**, 184-192.
- KORU-SENGUL, T., CLARK, J.D., FLEMING, L.E. & LEE, D.J. (2011). Toward improved statistical methods for analyzing Cotinine-Biomarker health association data. *Tobacco Induced Diseases* **9**, 11.
- LAI, T.L. & YING, Z. (1991). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *Annals of Statistics* **19**, 1370-1402.
- LIN, D. Y., ROBINS, J.M. & WEI, L.J. (1996). Comparing two failure time distributions in the presence of dependent censoring. *Biometrika* **83**, 381-393.
- LIN, D. Y., WEI, L.J. & YING, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557-572.
- LITTLE, R. J & RUBIN, D. B. (2002). Statistical Analysis with Missing Data. New York: Wiley.
- LU, X., NAN, B., SONG P. & SOWERS, M. (2010). Longitudinal data analysis with event time as a covariate. *Stat. Biosci.* **2**, 65-80.
- LUBIN, J.H., COLT, J.S. & CAMANN, D. (2004). Epidemiologic evaluation of measurement data in the presence of detection limits. *Environ Health Perspect* **112**, 1691-1696.
- LYNN, H. (2001). Maximum likelihood inference for left-censored HIV RNA data. *Statistics in medicine* **20**, 33-45.
- MAY, R.C., IBRAHIM, J.G. & CHU, H. (2011). Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. *Statistics in Medicine* **30**, 2551-2561.
- MCCULLAGH, P. & NELDER, J. (1989). Generalized Linear Models. Chapman and Hall/CRC.

- MOULTON, L.H., CURRIERO, F.C. & BARROSO, P.F. (2002). Mixture models for quantitative HIV RNA data. *Stat Meth Med Res.* **11**, 317-325.
- MURPHY, L.E., GOLLENBERG, A.L., BUCK LOUIS, G.M., KOSTYNIK, P.J. & SUNDARAM, R. (2010). Maternal Serum Preconception Polychlorinated Biphenyl Concentrations and Infant Birth Weight. *Environ Health Perspect* **118**, 297-302.
- NAN, B., KALBFLEISCH, J.D. & YU, M. (2009). Asymptotic theory for the semiparametric accelerated failure time model with missing data. *Annals of Statistics* **37**, 2351-2376.
- NAN, B. & WELLNER J. A. (2013). A general semiparametric Z-estimation approach for case-cohort studies. *Statistica Sinica* **23**, 1155-1180.
- NAVAS-ACIEN A, SILBERGELD EK, PASTOR-BARRIUSO R & GUALLAR E. (2008). Arsenic exposure and prevalence of type 2 diabetes in US adults. *Journal of American Medical Association.* **300**, 814-22.
- NIE, L, CHU, H, LIU, C, COLE, SR, VEXLER, A & SCHISTERMAN, EF. (2010). Linear regression with an independent variable subject to a detection limit. *Epidemiology* **21**, S17- S24.
- PENG, L. & FINE, J. P. (2006). Rank estimation of accelerated lifetime models with dependent censoring. *Journal of the American Statistical Association* **101**, 1085-1093.
- PRENTICE, R.L. & CAI, J. (1992). Covariance and survivor function estimation using censored multivariate time data. *Biometrika* **79**, 495-512.
- PRENTICE, R.L. & MOODIE, F.Z. (2004). Hazard-based nonparametric survivor function estimation. *J.R.Statist. Soc. B.* **66**, 305-319.
- RICHARDSON, D.B. & CIAMPI, A. (2003). Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *Am J Epidemiol.* **157**, 355-363.
- SCHISTERMAN, E.F. & LITTLE, R. J. (2010). Opening the black box of biomarker measurement error. *Epidemiology* **21**, S1- S3.
- SCHISTERMAN, E.F. VEXLER A., WHITCOMB B.W. & LIU A. (2006). The limitations due to exposure detection limits for regression models. *Am J Epidemiol.* **163**, 374-383.
- SOWERS, M.R., EYVAZZADEH, A.D., MCCONNELL, D., YOSEF, M., JANNAUSCH, M. L., ZHANG, D., HARLOW, S. AND RANDOLPH & J.F. JR. (2008). Anti-mullerian hormone and inhibin B in the definition of ovarian aging and the menopause transition. *J Clin Endocrinol Metab.* **93**, 3478-3483.

- SOWERS, M.R., MCCONNELL, D., YOSEF, M., JANNAUSCH, M. L., HARLOW, S. & RANDOLPH, J.F. JR. (2010). Relating smoking, obesity, insulin resistance, and ovarian biomarker changes to the final menstrual period. *Ann. N.Y. Acad. Sci.* **1204**, 95-103.
- THOMPSONAND, M. & NELSON, K.P. (2003). Linear regression with type I interval- and left-censored response data. *Annals of Statistics*. **18**, 354- 372.
- TSIATIS, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Annals of Statistics*. **18**, 354- 372.
- VAN DER LAAN, M.J. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE. *Annals of Statistics*. **24**, 596-627.
- VAN DER VAART, A. W. & WELLNER, J.A. (1996). Weak Convergence and Empirical Process. *Springer*.
- WEI, L. J. , YING, Z. & LIN, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika*. **77**, 845-851.
- WU, H., CHEN, Q., WARE, L.B. & KOYAMA, T. (2012). A Bayesian approach for generalized linear models with explanatory biomarker measurement variables subject to detection limit: an application to acute lung injury. *J. Appl Stat* **39**, 1733-1747.
- YU, M. & NAN, B. (2006). A hybrid Newton-type method for censored survival data using double weights in linear models. *Lifetime Data Anal* **12**, 345-364.
- YU, M. & NAN, B. (2009). Regression calibration in semiparametric accelerated failure time models. *Biometrics* **66**, 405- 414.
- ZHOU, M. (2006). The rankreg package.
<http://cran.r-project.org/src/contrib/Descriptions/rankreg.html>

Semiparametric Approach for Regression with Covariate Subject to Limit of Detection (Supplementary Data)

Shengchun Kong and Bin Nan

December 9, 2014

1 General Z-estimation theory

The proof of Theorem 1 in the main text is based on the general Z-estimation theory of Nan and Wellner (2013), which is provided in the following Lemmas 1.1 and 1.2 for our problem setting. Detailed discussion and proofs of these two lemmas can be found in Nan and Wellner (2013). Let $|\cdot|$ be the Euclidian norm and $\|\eta - \eta_0\| = \sup_t |\eta(t) - \eta_0(t)|$. Define $\rho\{(\phi, \alpha, \eta), (\phi_0, \alpha_0, \eta_0)\} = |\phi - \phi_0| + |\alpha - \alpha_0| + \|\eta - \eta_0\|$. We use P^* to denote outer probability, which is defined as $P^*(A) = \inf\{pr(B) : B \supset A, B \in \mathcal{B}\}$ for any subset A of Ω in a probability space (Ω, \mathcal{B}, P) .

Lemma 1.1. (*Consistency.*) Suppose θ_0 is the unique solution to $\Psi_\theta(\phi_0, \alpha_0, \eta_0) = 0$ in the parameter space Θ and $(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n})$ are estimators of $(\phi_0, \alpha_0, \eta_0)$ such that $\rho\{(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n}), (\phi_0, \alpha_0, \eta_0)\} = o_{P^*}(1)$. If

$$\sup_{\theta \in \Theta, \rho\{(\phi, \alpha, \eta), (\phi_0, \alpha_0, \eta_0)\} \leq \delta_n} \frac{|\Psi_{n, \theta}(\phi, \alpha, \eta) - \Psi_\theta(\phi_0, \alpha_0, \eta_0)|}{1 + |\Psi_{n, \theta}(\phi, \alpha, \eta)| + |\Psi_\theta(\phi_0, \alpha_0, \eta_0)|} = o_{P^*}(1) \quad (1)$$

for every sequence $\{\delta_n \downarrow 0\}$, then $\hat{\theta}_n$ satisfying $\Psi_{n, \hat{\theta}_n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n}) = o_{P^*}(1)$ converges in outer probability to θ_0 .

Lemma 1.2. (*Rate of convergence and asymptotic representation.*) Suppose that $\hat{\theta}_n$ satisfying $\Psi_{n, \hat{\theta}_n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n}) = o_{P^*}(n^{-1/2})$ is a consistent estimator of θ_0 that is a solution to $\Psi_\theta(\phi_0, \alpha_0, \eta_0) = 0$ in Θ , and that $(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n})$ is an estimator of $(\phi_0, \alpha_0, \eta_0)$ satisfying $\rho\{(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n}), (\phi_0, \alpha_0, \eta_0)\} = O_{P^*}(n^{-1/2})$. Suppose the following four conditions are satisfied:

(i) (*Stochastic equicontinuity.*)

$$\frac{|n^{1/2}(\Psi_{n, \hat{\theta}_n} - \Psi_{\hat{\theta}_n})(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n}) - n^{1/2}(\Psi_{n, \theta_0} - \Psi_{\theta_0})(\phi_0, \alpha_0, \eta_0)|}{1 + n^{1/2}|\Psi_{n, \hat{\theta}_n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n})| + n^{1/2}|\Psi_{\hat{\theta}_n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n})|} = o_{P^*}(1).$$

$$(ii) \ n^{1/2}\Psi_{n,\theta_0}(\phi_0, \alpha_0, \eta_0) = O_{p^*}(1).$$

(iii) (Smoothness.) There exist continuous matrices $\dot{\Psi}_{1,\theta_0}(\phi_0, \alpha_0, \eta_0)$, $\dot{\Psi}_{2,\theta_0}(\phi_0, \alpha_0, \eta_0)$, $\dot{\Psi}_{3,\theta_0}(\phi_0, \alpha_0, \eta_0)$, and a continuous linear functional $\dot{\Psi}_{4,\theta_0}(\phi_0, \alpha_0, \eta_0)$ such that

$$\begin{aligned} & |\Psi_{\hat{\theta}_n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}) - \Psi_{\theta_0}(\phi_0, \alpha_0, \eta_0) \\ & - \dot{\Psi}_{1,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\theta}_n - \theta_0) - \dot{\Psi}_{2,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\phi}_n - \phi_0) \\ & - \dot{\Psi}_{3,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\alpha}_n - \alpha_0) - \dot{\Psi}_{4,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\eta}_{n,\hat{\alpha}_n} - \eta_0)| \\ & = o(|\hat{\theta}_n - \theta_0|) + o[\rho\{(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}), (\phi_0, \alpha_0, \eta_0)\}]. \end{aligned} \quad (2)$$

Here the subscripts 1, 2, 3, and 4 correspond to θ , ϕ , α , and η in $\Psi_\theta(\phi, \alpha, \eta)$, respectively, and we assume that the matrix $\dot{\Psi}_{1,\theta_0}(\phi_0, \alpha_0, \eta_0)$ is nonsingular.

(iv) $n^{1/2}\dot{\Psi}_{2,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\phi}_n - \phi_0) = O_{p^*}(1)$, $n^{1/2}\dot{\Psi}_{3,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\alpha}_n - \alpha_0) = O_{p^*}(1)$, and $n^{1/2}\dot{\Psi}_{4,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\eta}_{n,\hat{\alpha}_n} - \eta_0) = O_{p^*}(1)$.

Then $\hat{\theta}_n$ is $n^{1/2}$ -consistent and further we have

$$\begin{aligned} n^{1/2}(\hat{\theta}_n - \theta_0) &= \{-\dot{\Psi}_{1,\theta_0}(\phi_0, \alpha_0, \eta_0)\}^{-1} n^{1/2}\{(\Psi_{n,\theta_0} - \Psi_{\theta_0})(\phi_0, \alpha_0, \eta_0) \\ &+ \dot{\Psi}_{2,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\phi}_n - \phi_0) + \dot{\Psi}_{3,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\alpha}_n - \alpha_0) \\ &+ \dot{\Psi}_{4,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\eta}_{n,\hat{\alpha}_n} - \eta_0)\} + o_{p^*}(1). \end{aligned} \quad (3)$$

2 Technical lemmas

Now we provide technical preparations for the proof of Theorem 1, some of which are from Ying Ding's 2010 University of Michigan Ph.D. thesis. We adopt the empirical process notation of van der vaart and Wellner (1996).

Let $\epsilon_\alpha = V - X'\alpha$ and $\epsilon_0 = V - X'\alpha_0$. Define

$$\begin{aligned} h^{(0)}(\alpha, s) &= P\{1(\epsilon_\alpha \leq s, \Delta = 1)\}, \\ h^{(1)}(\alpha, s) &= P\{1(\epsilon_\alpha \geq s)\}, \\ h^{(2)}(\alpha, s) &= P\{1(\epsilon_\alpha \geq s)X\}, \end{aligned}$$

and

$$H_n^{(1)}(\alpha, s) = \mathbb{P}_n\{1(\epsilon_\alpha \geq s)\}.$$

The Kaplan-Meier estimator of the distribution function of $T - \alpha X$ is given by

$$\hat{\eta}_{n,\alpha}(t) = 1 - \prod_{i: V_i - X'_i \alpha \leq t} \left\{ 1 - \frac{\Delta_i/n}{H_n^{(1)}(\alpha, V_i - X'_i \alpha)} \right\}.$$

Define

$$F(\alpha, t) = 1 - \exp \left\{ - \int_{u \leq t} \frac{dh^{(0)}(\alpha, u)}{h^{(1)}(\alpha, u)} \right\},$$

and denote $\dot{F}_\alpha(\alpha, t) = \partial F(\alpha, t)/\partial \alpha$. For function c in the exponential family, denote $\dot{c}_\phi(Y, \phi_0) = \partial c(Y, \phi)/\partial \phi|_{\phi=\phi_0}$.

Let $\Phi\{\alpha, h^{(1)}, h^{(2)}\} = P[\{h^{(1)}(\alpha, \epsilon_\alpha)X - h^{(2)}(\alpha, \epsilon_\alpha)\} \Delta]$, which corresponds to the limiting Gehan weighted estimating function, and define

$$\begin{aligned} m_1(\alpha_0, s; t) &= -P\left\{\frac{\Delta 1(s \geq \epsilon_0)1(t \geq \epsilon_0)}{h^{(1)}(\alpha_0, \epsilon_0)^2}\right\}, \quad m_2(\alpha_0, s; t, \Delta) = \frac{\Delta 1(t \geq s)}{h^{(1)}(\alpha_0, s)} \\ m_3(\alpha_0, \epsilon_0; \Delta, X) &= \left[-\dot{\Phi}_\alpha\{\alpha_0, h^{(1)}(\alpha_0, \cdot), h^{(2)}(\alpha_0, \cdot)\}\right]^{-1} \left[\{h^{(1)}(\alpha_0, \cdot)X - h^{(2)}(\alpha_0, \cdot)\} \Delta\right] \\ &\quad - \int \{1(\epsilon_0 \geq t)X\} dP_{\epsilon_0, \Delta}(t, 1) + \int \{1(\epsilon_0 \geq t)\} x dP_{\epsilon_0, \Delta, X}(t, 1, x). \end{aligned} \quad (4)$$

Lemma 2.1. *Suppose Conditions 3-4 hold, and let $\hat{\alpha}_n$ be the Gehan weighted estimator for α_0 , we have*

$$\sup_{t \in [C - E_5, \tau]} |\hat{\eta}_{n, \hat{\alpha}_n}(t) - \eta_0(t)| = O_{p^*}(n^{-1/2}),$$

where C is transformed L and $E_5 = \sup_{\alpha \in \mathcal{A}, x \in \mathcal{X}} |x'\alpha| < \infty$.

Proof. From the proof of Ying Ding's Theorem 2.2.3 in her 2010 University of Michigan Ph.D. thesis, for t in a bounded interval, we have for $t \in [C - E_5, \tau]$,

$$\begin{aligned} \sup_t n^{1/2} \{\hat{\eta}_{n, \hat{\alpha}_n}(t) - \eta_0(t)\} &= \sup_t \mathbb{G}_n[\{1 - \eta_0(t)\}\{m_1(\alpha_0, \epsilon_0; t) + m_2(\alpha_0, \epsilon_0; t, \Delta)\} \\ &\quad + \dot{F}_\alpha(\alpha_0, t)m_3(\alpha_0, \epsilon_0; X, \Delta)] + o_p(1), \end{aligned} \quad (6)$$

where $m_1(\alpha_0, s; t)$, $m_2(\alpha_0, \epsilon_0; t, \Delta)$, $m_3(\alpha_0, \epsilon_0; X, \Delta)$ are defined in (4) and (5).

We first calculate the bracket numbers for $\mathcal{F}_1 = \{m_1(\alpha_0, \epsilon_0; t), t \in [C - E_5, \tau]\}$ and $\mathcal{F}_2 = \{m_2(\alpha_0, \epsilon_0; t, \Delta), t \in [C - E_5, \tau]\}$. For any nontrivial ε satisfying $1 > \varepsilon > 0$, let t_i be the i -th $\lceil 1/\varepsilon \rceil$ quantile of $\varsigma_0 = T - X'\alpha_0$, i.e.

$$pr(\varsigma_0 \leq t_i) = i\varepsilon, \quad i = 1, \dots, \lceil 1/\varepsilon \rceil - 1,$$

where $\lceil x \rceil$ is the smallest integer that is greater than or equal to x . Furthermore, denote $t_0 = 0$ and $t_{\lceil 1/\varepsilon \rceil} = +\infty$. For $i = 1, \dots, \lceil 1/\varepsilon \rceil$, define brackets $[L_i, U_i]$ with

$$L_i(s) = -P\left\{\frac{\Delta 1(s \geq \epsilon_0)1(t_i \geq \epsilon_0)}{h^{(1)}(\alpha_0, \epsilon_0)^2}\right\}, \quad U_i(s) = -P\left\{\frac{\Delta 1(s \geq \epsilon_0)1(t_{i-1} \geq \epsilon_0)}{h^{(1)}(\alpha_0, \epsilon_0)^2}\right\}$$

such that $L_i(s) \leq -P\left\{\frac{\Delta 1(s \geq \epsilon_0)1(t \geq \epsilon_0)}{h^{(1)}(\alpha_0, \epsilon_0)^2}\right\} \leq U_i(s)$ when $t_{i-1} < t \leq t_i$. Since

$$E|U_i - L_i| \leq pr(t_{i-1} < \varsigma_0 \leq t_i)/\{h^{(1)}(\alpha_0, \tau)\}^2 = \varepsilon/\xi^2$$

from Condition 4, we have $N_{[\cdot]}(\varepsilon/\xi^2, \mathcal{F}_1, L_1) \leq 2/\varepsilon$ which yields

$$N_{[\cdot]}(\varepsilon, \mathcal{F}_1, L_1) \leq K_1/\varepsilon,$$

where $K_1 = 2\xi^2$. Similarly, we have

$$N_{[\cdot]}(\varepsilon, \mathcal{F}_2, L_1) \leq K_2/\varepsilon,$$

where $K_2 = 2\xi$. From Theorem 2.14.9 in van der vaart and Wellner (1996), we have

$$\begin{aligned} P^* \left(\sup_{t \in [C-E_5, \tau]} |\mathbb{G}_n \{(1 - \eta_0(t))m_1(\alpha_0, \epsilon_0; t)\}| > q \right) \\ \leq P^* \left(\sup_{t \in [C-E_5, \tau]} |\mathbb{G}_n \{m_1(\alpha_0, \epsilon_0; t)\}| > q \right) \leq D_1 q e^{-2q^2}, \end{aligned} \quad (7)$$

$$\begin{aligned} P^* \left(\sup_{t \in [C-E_5, \tau]} |\mathbb{G}_n \{(1 - \eta_0(t))m_2(\alpha_0, \epsilon_0; t, \Delta)\}| > q \right) \\ \leq P^* \left(\sup_{t \in [C-E_5, \tau]} |\mathbb{G}_n \{m_2(\alpha_0, \epsilon_0; t, \Delta)\}| > q \right) \leq D_2 q e^{-2q^2} \end{aligned} \quad (8)$$

for some constant D_1 depends on K_1 and constant D_2 depends on K_2 . We now show $\sup_{t \in [C-E_5, \tau]} |\dot{F}_\alpha(\alpha_0, t)|$ is bounded. Direct calculation yields

$$\begin{aligned} & \sup_{t \in [C-E_5, \tau]} |\dot{F}_\alpha(\alpha_0, t)| \\ &= \sup_{t \in [C-E_5, \tau]} e^{-\int_{u \leq t} \frac{dh^{(0)}(\alpha_0, u)}{h^{(1)}(\alpha_0, u)}} \left| \int_{u \leq t} \frac{d\dot{h}_\alpha^{(0)}(\alpha_0, u)}{h^{(1)}(\alpha_0, u)} - \int_{u \leq t} \frac{\dot{h}_\alpha^{(1)}(\alpha_0, u) dh^{(0)}(\alpha_0, u)}{\{h^{(1)}(\alpha_0, u)\}^2} \right| \\ &\leq \{h^{(1)}(\alpha_0, \tau)\}^{-1} \sup_{t \in [C-E_5, \tau]} |\dot{h}_\alpha^{(0)}(\alpha_0, t)| \\ &\quad + \{h^{(1)}(\alpha_0, \tau)\}^{-2} \sup_{u \in (-\infty, \infty)} \left| \dot{h}_u^{(0)}(\alpha_0, u) \right| \sup_{t \in [C-E_5, \tau]} \int_{u \leq t} |\dot{h}_\alpha^{(1)}(\alpha_0, u)| du, \end{aligned}$$

where $\dot{h}_\alpha^{(0)}(\alpha_0, t) = \frac{\partial}{\partial \alpha} h^{(0)}(\alpha, t)|_{\alpha=\alpha_0}$, $\dot{h}_\alpha^{(1)}(\alpha_0, t) = \frac{\partial}{\partial \alpha} h^{(1)}(\alpha, t)|_{\alpha=\alpha_0}$ and $\dot{h}_u^{(0)}(\alpha_0, u) = \frac{\partial}{\partial u} h^{(0)}(\alpha_0, u)$. Since

$$\begin{aligned} h^{(0)}(\alpha, t) &= \int \eta_0(\min(t + x'\alpha - x'\alpha_0, C - x'\alpha_0)) dF_X(x) \\ &= \int_{x'\alpha \geq C-t} \eta_0(C - x'\alpha_0) dF_X(x) + \int_{x'\alpha < C-t} \eta_0(t + x'\alpha - x'\alpha_0) dF_X(x), \\ h^{(1)}(\alpha, t) &= \int_{x'\alpha \leq C-t} \{1 - \eta_0(t + x'\alpha - x'\alpha_0)\} dF_X(x), \end{aligned}$$

where $F_X(x)$ is the distribution function of X , from Condition 3 we have

$$\begin{aligned}
\sup_{t \in [C-E_5, \tau]} |\dot{h}_\alpha^{(0)}(\alpha_0, t)| &= \sup_{t \in [C-E_5, \tau]} \left| \dot{\eta}_{0, \alpha_0}(t) \int_{t+x'\alpha_0 < C} x dF_X(x) \right| \leq E_3 E|X| < \infty, \\
\sup_{u \in (-\infty, \infty)} |\dot{h}_u^{(0)}(\alpha_0, u)| &\leq \sup_{u \in (-\infty, \infty)} |\dot{\eta}_{0, \alpha_0}(u)| \leq E_3, \\
\sup_{t \in [C-E_5, \tau]} \int_{u \leq t} |\dot{h}_\alpha^{(1)}(\alpha_0, u)| du &\leq \sup_{t \in [C-E_5, \tau]} \int_{u \leq t} \left| \int_{t+x'\alpha_0 \leq C} x dF_X(x) \right| \dot{\eta}_{0, \alpha_0}(u) du + \int_{-\infty}^{\infty} |x| dF_X(x) \\
&\leq E|X|E_3 + E|X| < \infty.
\end{aligned}$$

Since it can be shown that $m_3(\alpha_0, \epsilon_0; X, \Delta)$ has finite second moment, we have $\sup_{t \in [C-E_5, \tau]} \mathbb{G}_n[\dot{F}_\alpha(\alpha_0, t)m_3(\alpha_0, \epsilon_0; X, \Delta)] = O_{p^*}(1)$, thus obtain the desired result. \square

Lemma 2.2. *Suppose Condition 7 holds, we have that*

$$\left\{ \Delta\{Y - \dot{b}(D'(t)\theta)\}D(t), \theta \in \Theta, t \in \mathcal{T} \subset \mathbb{R} \right\} \quad (9)$$

is Donsker.

Proof. From Condition 7 we know that $\ddot{b}(\cdot)$ is bounded, hence $\dot{b}(\cdot)$ is a Lipschitz function. From Theorem 2.10.6 in van der vaart and Wellner (1996), we know that $D(t)$ and $\dot{b}(D'(t)\theta)$ are Donsker, hence (9) is Donsker. \square

Lemma 2.3. *Suppose \mathcal{X} and \mathcal{A} be the bounded covariate and parameter spaces. Let \mathcal{H} be a collection of distribution functions satisfying Condition 3. We have $\mathcal{F} = \{\eta(t - x'\alpha), t \in \mathcal{T} \subset \mathbb{R}, x \in \mathcal{X}, \alpha \in \mathcal{A}, \eta \in \mathcal{H}\}$ is Donsker.*

Proof. Let $\mathcal{F}_1 = \{\eta(t)\}$. From Theorem 2.7.5 in van der vaart and Wellner (1996), the number of brackets $[L_i, U_i]$ such that $L_i(t) \leq \eta(t) \leq U_i(t)$ for any nontrivial ε with $1 > \varepsilon > 0$ and $\int |U_i(t) - L_i(t)| d\eta_0(t) \leq \varepsilon$ satisfies $\log N_{[\cdot]}(\varepsilon, \mathcal{F}_1, L_1(P)) \leq K_1/\varepsilon$, where $K_1 < \infty$ is a constant.

For notational simplicity, we consider 1-dimensional \mathcal{A} . Because \mathcal{A} is bounded, we partition \mathcal{A} by a set of intervals $[l_k, u_k]$ such that $|u_k - l_k| \leq \varepsilon$. Hence the number of such intervals is bounded by K_2/ε with a constant $K_2 < \infty$. Now we construct brackets for $\mathcal{F} \equiv \{\eta(t - x\alpha)\}$. Define

$$O_{ik}(t, x) = \min(L_i(t - xu_k), L_i(t - xl_k)), S_{ik}(t, x) = \max(U_i(t - xu_k), U_i(t - xl_k)).$$

We have

$$\begin{aligned}
O_{ik}(t, x) &\leq \min(\eta(t - xl_k), \eta(t - xu_k)) \\
&\leq \eta(t - x\alpha) \\
&\leq \max(\eta(t - xl_k), \eta(t - xu_k)) \leq S_{ik}(t, x).
\end{aligned}$$

Since

$$P | S_{ik} - O_{ik} | \leq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} | U_i(t - xu_k) - L_i(t - xu_k) | d\eta_0(t + x\alpha_0) dF_X(x) \quad (10)$$

$$+ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} | U_i(t - xl_k) - L_i(t - xl_k) | d\eta_0(t + x\alpha_0) dF_X(x) \quad (11)$$

$$+ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} | U_i(t - xl_k) - L_i(t - xu_k) | d\eta_0(t + x\alpha_0) dF_X(x) \quad (12)$$

$$+ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} | U_i(t - xu_k) - L_i(t - xl_k) | d\eta_0(t + x\alpha_0) dF_X(x). \quad (13)$$

Since $[L_i, U_i]$ are brackets for \mathcal{F}_1 , we have (10) $\leq \varepsilon$ and (11) $\leq \varepsilon$. Furthermore, by integration by parts and change of variables we obtain

$$\begin{aligned} (12) &\leq 2\varepsilon + \int_0^{\infty} \int_{-\infty}^{\infty} \{\eta(t - xl_k) - \eta(t - xu_k)\} d\eta_0(t + x\alpha_0) dF_X(x) \\ &\quad + \int_{-\infty}^0 \int_{-\infty}^{\infty} \{\eta(t - xu_k) - \eta(t - xl_k)\} d\eta_0(t + x\alpha_0) dF_X(x) \\ &= 2\varepsilon + \int_0^{\infty} \int_{-\infty}^{\infty} \{\eta_0(t + x\alpha_0 + xu_k) - \eta_0(t + x\alpha_0 + xl_k)\} d\eta(t) dF_X(x) \\ &\quad + \int_{-\infty}^0 \int_{-\infty}^{\infty} \{\eta_0(t + x\alpha_0 + xl_k) - \eta_0(t + x\alpha_0 + xu_k)\} d\eta(t) dF_X(x) \\ &\leq 2\varepsilon + \int_0^{\infty} \int_{-\infty}^{\infty} E_3 x(u_k - l_k) d\eta(t) dF_X(x) - \int_{-\infty}^0 \int_{-\infty}^{\infty} E_3 x(u_k - l_k) d\eta(t) dF_X(x) \\ &\leq 2\varepsilon + E_3 E|X| \varepsilon = K_3 \varepsilon, \end{aligned}$$

where E_3 is defined in Condition 3, and $K_3 = 2 + E_3 E|X| < \infty$. Similarly, we have (13) $\leq K_3 \varepsilon$. Hence we have $N_{[\cdot]}((2 + 2K_3)\varepsilon, \mathcal{F}, L_1(P)) \leq \exp(K_1/\varepsilon) K_2/\varepsilon$, i.e. $N_{[\cdot]}(\varepsilon, \mathcal{F}, L_1(P)) \leq \exp(K_1(2 + 2K_3)/\varepsilon) K_2(2 + 2K_3)/\varepsilon \leq \exp((K_1 + K_2)(2 + 2K_3)/\varepsilon)$. Hence, \mathcal{F} is Donsker. \square

Lemma 2.4. *Suppose Conditions 2, 5-9 hold, we have*

$$\left\{ \frac{\int_{C-x'\alpha}^{\tau} f_{\theta, \phi}(y | t + x'\alpha, x) \{y - \dot{b}(D'(t + x'\alpha)\theta)\} D(t + x'\alpha) d\eta(t)}{\int_{C-x'\alpha}^{\tau} f_{\theta, \phi}(y | t + x'\alpha, x) d\eta(t)} : \right. \quad (14)$$

$$\left. \theta \in \Theta, |1/a(\phi)| < l, \alpha \in \mathcal{A}, \eta \in \mathcal{H}, \rho\{(\phi, \alpha, \eta), (\phi_0, \alpha_0, \eta_0)\} < \delta_2, x \in \mathcal{X}, y \in \mathcal{Y} \right\}$$

is Donsker.

Proof. From Condition 9, we have $\{\int_{C-x'\alpha}^{\tau} f_{\theta, \phi}(y | t + x'\alpha, x) d\eta(t)\}$ bounded away from zero. From Section 2.10.2 of van der vaart and Wellner (1996), we

only need to show that both the numerator and denominator in (14) belong to Donsker classes. By integration by parts, we have

$$\begin{aligned} & \int_{C-x'\alpha}^{\tau} f_{\theta,\phi}(y \mid t+x'\alpha, x) d\eta(t) \\ &= f_{\theta,\phi}(y \mid \tau+x'\alpha, x)\eta(\tau) - f_{\theta,\phi}(y \mid C, x)\eta(C-x'\alpha) \\ & \quad - \int_{C-E_5}^{\tau} 1(t \geq C-x'\alpha)\eta(t)f_{\theta,\phi}(y \mid t+x'\alpha, x) \\ & \quad \gamma\{y - \dot{b}(D'(t+x'\alpha)\theta)\}\dot{h}(t+x'\alpha)/a(\phi)dt. \end{aligned}$$

In the above, $\dot{h}(\cdot)$ is Lipschitz by Condition 2 and $f_{\theta,\phi}(y \mid t+x'\alpha, x)$ is Lipschitz function for θ , ϕ and α by Conditions 2, 5 and 8, thus both belong to Donsker classes by Theorem 2.10.6 in van der vaart and Wellner (1996). By Lemma 2.3 we know that $\{\eta(C-x'\alpha)\}$ is Donsker. Since the class of indicator functions of half spaces is a VC-class, see e.g. Exercise 9 on page 151 and Exercise 14 on page 152 in van der vaart and Wellner (1996), thus the set of functions $\{1(t \geq C-x'\alpha)\}$ is a Donsker class. By Theorem 2.10.3 in van der vaart and Wellner (1996), the permanence of the Donsker property for the closure of the convex hull, we have $\left\{ \int_{C-E_5}^{\tau} 1(t \geq C-x'\alpha)\eta(t)f_{\theta,\phi}(y \mid t+x'\alpha, x)\gamma\{y - \dot{b}(D'(t+x'\alpha)\theta)\}\dot{h}(t+x'\alpha)/a(\phi)dt \right\}$ is Donsker. Hence the denominator of (14) belongs to a Donsker class.

Similarly, by integration by parts,

$$\begin{aligned} & \int_{C-x'\alpha}^{\tau} f_{\theta,\phi}(y \mid t+x'\alpha, x)\{y - \dot{b}(D'(t+x'\alpha)\theta)\}D(t+x'\alpha)d\eta(t) \\ &= f_{\theta,\phi}(y \mid \tau+x'\alpha, x)\{y - \dot{b}(D'(\tau+x'\alpha)\theta)\}D(\tau+x'\alpha)\eta(\tau) \\ & \quad - f_{\theta,\phi}(y \mid C, x)\{y - \dot{b}(D'(C)\theta)\}D(C)\eta(C-x'\alpha) \\ & \quad - \int_{C-E_5}^{\tau} 1(t \geq C-x'\alpha)\eta(t)f_{\theta,\phi}(y \mid t+x'\alpha, x) \\ & \quad (\gamma\{y - \dot{b}(D'(t+x'\alpha)\theta)\}^2 D(t+x'\alpha)/a(\phi) - \ddot{b}(D'(t+x'\alpha)\theta)\gamma D(t+x'\alpha) \\ & \quad + \{y - \dot{b}(D'(t+x'\alpha)\theta)\}J_{p+2})\dot{h}(t+x'\alpha)dt, \end{aligned}$$

where $J_{p+2} = (0, \dots, 0, 1)'_{1 \times (p+2)}$. Similar to the denominator, we can show that the above function, which is the numerator of (14), belongs to a Donsker class provided that $\{\ddot{b}(D'(t+x'\alpha)\theta)\}$ is Donsker from Condition 7. \square

Lemma 2.5. *Under Conditions 5-9, when $\theta \rightarrow \theta_0$ and $\rho\{(\phi, \alpha, \eta), (\phi_0, \alpha_0, \eta_0)\} \rightarrow 0$, we have that $E|\psi_{\theta}(\phi, \alpha, \eta) - \psi_{\theta_0}(\phi_0, \alpha_0, \eta_0)|^2 \rightarrow 0$.*

Proof. The proof follows straightforward algebraic calculations based on the Mean Value Theorem. The details are thus omitted. \square

Lemma 2.6. *Suppose Conditions 2, 5-9 hold, we have $E|\psi_{\theta_0}(\phi_0, \alpha_0, \eta_0)|^2 < \infty$.*

Proof. Again, the proof is based on direct calculation. \square

3 Proof of Theorem 1

3.1 Proof of consistency

Proof. We prove consistency using Lemma 1.1. Since $\hat{\phi}_n$ and $\hat{\alpha}_n$ are $n^{1/2}$ -consistent, see the last paragraph of Section 3, and $\hat{\eta}_{n,\hat{\alpha}_n}$ is also $n^{1/2}$ -consistent in a finite interval from Lemma 2.1, we have

$$\rho\{(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}), (\phi_0, \alpha_0, \eta_0)\} = o_{p^*}(1).$$

Given that θ_0 is the unique solution to $\Psi_\theta(\phi_0, \alpha_0, \eta_0) = 0$ from Condition 1, we only need to show that

$$\sup_{\theta \in \Theta, \rho\{(\phi, \alpha, \eta), (\phi_0, \alpha_0, \eta_0)\} \leq \delta_n} |\Psi_{\theta,n}(\phi, \alpha, \eta) - \Psi_\theta(\phi_0, \alpha_0, \eta_0)| = o_{p^*}(1) \quad (15)$$

for every sequence $\delta_n \downarrow 0$. Now

$$\begin{aligned} & \sup_{\theta \in \Theta, \rho\{(\phi, \alpha, \eta), (\phi_0, \alpha_0, \eta_0)\} \leq \delta_n} |\Psi_{\theta,n}(\phi, \alpha, \eta) - \Psi_\theta(\phi_0, \alpha_0, \eta_0)| \\ & \leq \sup_{\theta \in \Theta} |(\mathbb{P}_n - P)[\Delta\{Y - \dot{b}(D(T)\theta)\}D(T)]| \end{aligned} \quad (16)$$

$$+ \sup_{\theta \in \Theta, \rho\{(\phi, \alpha, \eta), (\phi_0, \alpha_0, \eta_0)\} \leq \delta_n} \quad (17)$$

$$\begin{aligned} & P \left| \frac{\int_{C-X'\alpha}^\tau f_{\theta,\phi}(Y | t + X'\alpha, X) \{Y - \dot{b}(D'(t + X'\alpha)\theta)\} D(t + X'\alpha) d\eta(t)}{\int_{C-X'\alpha}^\tau f_{\theta,\phi}(Y | t + X'\alpha, X) d\eta(t)} \right. \\ & \left. - \frac{\int_{C-X'\alpha_0}^\tau f_{\theta,\phi_0}(Y | t + X'\alpha_0, X) \{Y - \dot{b}(D'(t + X'\alpha_0)\theta)\} D(t + X'\alpha_0) d\eta_0(t)}{\int_{C-X'\alpha_0}^\tau f_{\theta,\phi_0}(Y | t + X'\alpha_0, X) d\eta_0(t)} \right| \\ & + \sup_{\theta \in \Theta, \rho\{(\phi, \alpha, \eta), (\phi_0, \alpha_0, \eta_0)\} \leq \delta_n} \left| (\mathbb{P}_n - P)(1 - \Delta) \right. \\ & \left. \frac{\int_{C-X'\alpha}^\tau f_{\theta,\phi}(Y | t + X'\alpha, X) \{Y - \dot{b}(D'(t + X'\alpha)\theta)\} D(t + X'\alpha) d\eta(t)}{\int_{C-X'\alpha}^\tau f_{\theta,\phi}(Y | t + X'\alpha, X) d\eta(t)} \right| = o_{p^*}(1), \end{aligned} \quad (18)$$

where (16) and (18) equal to $o_{p^*}(1)$ are from Lemma 2.2 and Lemma 2.4, respectively, and (17) equal to $o_{p^*}(1)$ follows a direct calculation similar to Lemma 2.5 using the Mean Value Theorem. \square

3.2 Proof of asymptotic normality

Proof. We now verify all the conditions in Lemma 1.2. Condition (i) holds because $\{\psi_\theta(\phi, \alpha, \eta)\}$ is Donsker by Lemmas 2.2 and 2.4, together with the result in Lemma 2.5. Condition (ii) holds by the classical central limit theorem for independent and identically distributed data with $E|\psi_{\theta_0}(\phi_0, \alpha_0, \eta_0(\alpha_0))|^2 < \infty$ from Lemma 2.6.

For Condition (iii), given that $\rho\{(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n}), (\phi_0, \alpha_0, \eta_0)\} = O_{p^*}(n^{-1/2})$, taking the Taylor expansion for θ , ϕ and α we obtain

$$\begin{aligned} & \Psi_{\hat{\theta}_n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n}) - \Psi_{\theta_0}(\phi_0, \alpha_0, \eta_0) \\ &= \dot{\Psi}_{1, \tilde{\theta}}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n})(\hat{\theta}_n - \theta_0) - \dot{\Psi}_{2, \theta_0}(\tilde{\phi}, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n})(\hat{\phi}_n - \phi_0) \\ & \quad - \dot{\Psi}_{3, \theta_0}(\phi_0, \tilde{\alpha}, \eta_0)(\hat{\alpha}_n - \alpha_0) - R(\theta_0, \phi_0, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n}, \eta_0), \end{aligned}$$

where $\tilde{\theta}$ is between θ_0 and $\hat{\theta}_n$, $\tilde{\phi}$ is between ϕ_0 and $\hat{\phi}_n$, $\tilde{\alpha}$ is between α_0 and $\hat{\alpha}_n$, and the remainder has the following form

$$\begin{aligned} & R(\theta_0, \phi_0, \alpha, \eta, \eta_0) \\ &= P \left[(1 - \Delta) \left\{ \frac{\int_{C-X'\alpha}^{\tau} A(t, \theta_0, \phi_0, \alpha) d\eta(t)}{\int_{C-X'\alpha}^{\tau} B(t, \theta_0, \phi_0, \alpha) d\eta(t)} - \frac{\int_{C-X'\alpha}^{\tau} A(t, \theta_0, \phi_0, \alpha) d\eta_0(t)}{\int_{C-X'\alpha}^{\tau} B(t, \theta_0, \phi_0, \alpha) d\eta_0(t)} \right\} \right] \end{aligned}$$

with

$$\begin{aligned} A(t, \theta_0, \phi_0, \alpha) &= f_{\theta_0, \phi_0}(Y | t + X'\alpha, X) \{Y - \dot{b}(D'(t + X'\alpha)\theta_0)\} D(t + X'\alpha), \\ B(t, \theta_0, \phi_0, \alpha) &= f_{\theta_0, \phi_0}(Y | t + X'\alpha, X). \end{aligned}$$

It can be show by direct calculation that $|\dot{\Psi}_{1, \tilde{\theta}}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n}) - \dot{\Psi}_{1, \theta_0}(\phi_0, \alpha_0, \eta_0)| = o_{p^*}(1)$, $|\dot{\Psi}_{2, \theta_0}(\tilde{\phi}, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n}) - \dot{\Psi}_{2, \theta_0}(\phi_0, \alpha_0, \eta_0)| = o_{p^*}(1)$ and $|\dot{\Psi}_{3, \theta_0}(\phi_0, \tilde{\alpha}, \eta_0) - \dot{\Psi}_{3, \theta_0}(\phi_0, \alpha_0, \eta_0)| = o_{p^*}(1)$.

Define

$$\begin{aligned} & \dot{\Psi}_{4, \theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\eta}_{n, \hat{\alpha}_n} - \eta_0) \tag{19} \\ &= P \left[(1 - \Delta) \left\{ \frac{\int_{C-X'\alpha_0}^{\tau} A(t, \theta_0, \phi_0, \alpha_0) d[\hat{\eta}_{n, \hat{\alpha}_n}(t) - \eta_0(t)]}{\int_{C-X'\alpha_0}^{\tau} B(t, \theta_0, \phi_0, \alpha_0) d\eta_0(t)} \right. \right. \\ & \quad \left. \left. - \frac{\int_{C-X'\alpha_0}^{\tau} A(t, \theta_0, \phi_0, \alpha_0) d\eta_0(t) \int_{C-X'\alpha_0}^{\tau} B(t, \theta_0, \phi_0, \alpha_0) d[\hat{\eta}_{n, \hat{\alpha}_n}(t) - \eta_0(t)]}{\int_{C-X'\alpha_0}^{\tau} B(t, \theta_0, \phi_0, \alpha_0) d\eta_0(t)^2} \right\} \right] \\ &= P \left[(1 - \Delta) \left\{ \frac{\int_{C-X'\alpha_0}^{\tau} A(t, \theta_0, \phi_0, \alpha_0) d\hat{\eta}_{n, \hat{\alpha}_n}(t)}{\int_{C-X'\alpha_0}^{\tau} B(t, \theta_0, \phi_0, \alpha_0) d\eta_0(t)} \right. \right. \\ & \quad \left. \left. - \frac{\int_{C-X'\alpha_0}^{\tau} A(t, \theta_0, \phi_0, \alpha_0) d\eta_0(t) \int_{C-X'\alpha_0}^{\tau} B(t, \theta_0, \phi_0, \alpha_0) d\hat{\eta}_{n, \hat{\alpha}_n}(t)}{\int_{C-X'\alpha_0}^{\tau} B(t, \theta_0, \phi_0, \alpha_0) d\eta_0(t)^2} \right\} \right]. \end{aligned}$$

Then we have

$$\begin{aligned} & |R(\theta_0, \phi_0, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n}, \eta_0) - \dot{\Psi}_{4, \theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\eta}_{n, \hat{\alpha}_n} - \eta_0)| \\ & \leq |R(\theta_0, \phi_0, \hat{\alpha}_n, \hat{\eta}_{n, \hat{\alpha}_n}, \eta_0) - R(\theta_0, \phi_0, \alpha_0, \hat{\eta}_{n, \hat{\alpha}_n}, \eta_0)| \\ & \quad + |R(\theta_0, \phi_0, \alpha_0, \hat{\eta}_{n, \hat{\alpha}_n}, \eta_0) - \dot{\Psi}_{4, \theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\eta}_{n, \hat{\alpha}_n} - \eta_0)| \\ & = D_1 + D_2. \end{aligned}$$

Now $D_1 = o(|\hat{\alpha}_n - \alpha_0| + \|\hat{\eta}_{n, \hat{\alpha}_n} - \eta_0\|)$ can be shown by

$$\begin{aligned} & \frac{A_1}{B_1} - \frac{A_2}{B_2} - \frac{A_3}{B_3} + \frac{A_4}{B_4} \\ &= \frac{A_1}{B_1 B_2} (B_2 - B_1 - B_4 + B_3) + \frac{A_1}{B_1 B_2 B_3 B_4} (B_3 B_4 - B_1 B_2) (B_4 - B_3) \\ & \quad + \frac{A_1 - A_3}{B_3 B_4} (B_4 - B_3) + \frac{A_1 - A_2}{B_2 B_4} (B_4 - B_2) + \frac{A_1 - A_2 - A_3 + A_4}{B_4}, \end{aligned}$$

and $D_2 = o(|\hat{\alpha}_n - \alpha_0| + \|\hat{\eta}_{n, \hat{\alpha}_n} - \eta_0\|)$ can be shown by

$$\frac{A_1}{B_1} - \frac{A_2}{B_2} - \frac{A_1}{B_2} + \frac{A_2 B_1}{B_2^2} = \frac{1}{B_1 B_2^2} \{A_1 (B_1 - B_2)^2 - B_1 (A_2 - A_1) (B_2 - B_1)\}.$$

Since $\hat{\phi}_n$, $\hat{\alpha}_n$ and $\hat{\eta}_n$ are all root- n consistent, under Conditions (i)-(iii), Condition (iv) holds automatically. Then by Lemma 1.2 we have that $\hat{\theta}_n$ is $n^{1/2}$ -consistent and (3) holds with

$$\begin{aligned} & \dot{\Psi}_{1, \theta_0}(\phi_0, \alpha_0, \eta_0) \\ &= E[\Delta \ddot{b}\{D'(T)\theta_0\} D(T) D'(T)] \\ & \quad - E\left[(1 - \Delta) \left\{ \int_{C - X' \alpha_0}^{\tau} f_{\theta_0, \phi_0}(Y | t + X' \alpha_0, X) d\eta_0(t) \right\}^{-2} \right. \\ & \quad \left. \left(\int_{C - X' \alpha_0}^{\tau} f_{\theta_0, \phi_0}(Y | t + X' \alpha_0, X) \{Y - \dot{b}(D'(t + X' \alpha_0)\theta_0)\} \right. \right. \\ & \quad \left. \left. D(t + X' \alpha_0) d\eta_0(t) \right)^{\otimes 2} \right], \end{aligned}$$

$$\begin{aligned} & \dot{\Psi}_{2, \theta_0}(\phi_0, \alpha_0, \eta_0) \\ &= -E\left[(1 - \Delta) \left\{ \int_{C - X' \alpha_0}^{\tau} f_{\theta_0, \phi_0}(Y | t + X' \alpha_0, X) d\eta_0(t) \right\}^{-1} \right. \\ & \quad \left\{ \int_{C - X' \alpha_0}^{\tau} f_{\theta_0, \phi_0}(Y | t + X' \alpha_0, X) \{Y - \dot{b}(D'(t + X' \alpha_0)\theta_0)\} D(t + X' \alpha_0) \right. \\ & \quad \left. \left([Y \{D'(t + X' \alpha_0)\theta_0\} - b(D'(t + X' \alpha_0)\theta_0)] a'(\phi_0)/a(\phi_0)^2 - \dot{c}_\phi(Y, \phi_0) \right) d\eta_0(t) \right\} \\ & \quad - \left\{ \int_{C - X' \alpha_0}^{\tau} f_{\theta_0, \phi_0}(Y | t + X' \alpha_0, X) d\eta_0(t) \right\}^{-2} \left\{ \int_{C - X' \alpha_0}^{\tau} f_{\theta_0, \phi_0}(Y | t + X' \alpha_0, X) \right. \\ & \quad \left. \left([Y \{D'(t + X' \alpha_0)\theta_0\} - b(D'(t + X' \alpha_0)\theta_0)] a'(\phi_0)/a(\phi_0)^2 - \dot{c}_\phi(Y, \phi_0) \right) d\eta_0(t) \right. \\ & \quad \left. \left. \int_{C - X' \alpha_0}^{\tau} f_{\theta_0, \phi_0}(Y | t + X' \alpha_0, X) \{Y - \dot{b}(D'(t + X' \alpha_0)\theta_0)\} D(t + X' \alpha_0) d\eta_0(t) \right\} \right], \end{aligned}$$

and

$$\begin{aligned}
& \dot{\Psi}_{3,\theta_0}(\phi_0, \alpha_0, \eta_0) \\
&= -E \left[(1 - \Delta) \left\{ \int_{C-X'\alpha_0}^{\tau} f_{\theta_0, \phi_0}(Y | t + X'\alpha_0, X) d\eta_0(t) \right\}^{-2} \right. \\
&\quad \int_{C-X'\alpha_0}^{\tau} f_{\theta_0, \phi_0}(Y | t + X'\alpha_0, X) \{Y - \dot{b}(D'(t + X'\alpha_0)\theta_0)\} D(t + X'\alpha_0) d\eta_0(t) \\
&\quad \left\{ \int_{C-X'\alpha_0}^{\tau} f_{\theta_0, \phi_0}(Y | t + X'\alpha_0, X) \{Y - \dot{b}(D'(t + X'\alpha_0)\theta_0)\} \right. \\
&\quad \left. \left. \gamma_0 X' \dot{h}(t + X'\alpha_0) / a(\phi_0) d\eta_0(t) + f_{\theta_0, \phi_0}(Y | C, X) \dot{\eta}_0(C - X'\alpha_0) X' \right\} \right].
\end{aligned}$$

Finally, we obtain

$$\begin{aligned}
n^{1/2} \{(\Psi_{n, \theta_0} - \Psi_{\theta_0})(\phi_0, \alpha_0, \eta_0)\} &= \mathbb{G}_n \left(\Delta \{Y - \dot{b}(D'(T)\theta_0)\} D(T) \right. \\
&\quad \left. + (1 - \Delta) \left\{ \int_{C-X'\alpha_0}^{\tau} f_{\theta_0, \phi_0}(Y | t + X'\alpha_0, X) d\eta_0(t) \right\}^{-1} \right. \\
&\quad \left. \int_{C-X'\alpha_0}^{\tau} f_{\theta_0, \phi_0}(Y | t + X'\alpha_0, X) \{Y - \dot{b}(D'(t + X'\alpha_0)\theta_0)\} D(t + X'\alpha_0) d\eta_0(t) \right) \\
&= \mathbb{G}_n \left\{ G_1(\theta_0, \phi_0, \alpha_0, \eta_0, \Delta, Y, X, V) \right\} \tag{20}
\end{aligned}$$

and

$$\begin{aligned}
& n^{1/2} \dot{\Psi}_{2, \theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\phi}_n - \phi_0) \\
&= \mathbb{G}_n \left\{ \dot{\Psi}_{2, \theta_0}(\phi_0, \alpha_0, \eta_0) m_4(\theta_0, \Delta, Y, X, V) \right\} + o_p(1), \tag{21}
\end{aligned}$$

where $n^{1/2}(\hat{\phi}_n - \phi_0) = \mathbb{G}_n m_4(\theta_0, \phi_0, Y, X) + o_p(1)$ with $m_4(\theta_0, \phi_0, Y, X) = \Delta \{Y - D'(T)\theta_0\}^2$ for linear regression and $m_4 = 0$ for the logistic and Poisson regressions. For Gehan weighted estimate $\hat{\alpha}_n$, we have

$$n^{1/2} \dot{\Psi}_{3, \theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\alpha}_n - \alpha_0) = \mathbb{G}_n \left\{ \dot{\Psi}_{3, \theta_0}(\phi_0, \alpha_0, \eta_0) m_3(\alpha_0, \epsilon_0; \Delta, X) \right\} + o_p(1). \tag{22}$$

Furthermore, from (6) and (19) we obtain

$$\begin{aligned}
& n^{1/2} \dot{\Psi}_{4,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\eta}_{n,\hat{\alpha}_n} - \eta_0) \\
&= \mathbb{G}_n \left[- \int_{\mathcal{X}} \int_{-\infty}^{\infty} (1 - \Delta) \left\{ f_{\theta_0, \phi_0}(y \mid \tau + x' \alpha_0, x) \left(\{1 - \eta_0(\tau)\} \{m_1(\alpha_0, \epsilon_0; \tau) \right. \right. \right. \\
&\quad \left. \left. \left. + m_2(\alpha_0, \epsilon_0; \tau, \Delta)\} + \dot{F}_\alpha(\alpha, \tau) m_3(\alpha_0, \epsilon_0; x, \Delta) \right) \right. \right. \\
&\quad \left. - f_{\theta_0, \phi_0}(y \mid C, x) \left(\{1 - \eta_0(C - x' \alpha_0)\} \{m_1(\alpha_0, \epsilon_0; C - x' \alpha_0) \right. \right. \\
&\quad \left. \left. + m_2(\alpha_0, \epsilon_0; C - x' \alpha_0, \Delta)\} + \dot{F}_\alpha(\alpha, C - x' \alpha_0) m_3(\alpha_0, \epsilon_0; x, \Delta) \right) \right. \\
&\quad \left. - \int_{C - x' \alpha_0}^{\tau} f_{\theta_0, \phi_0}(y \mid t + x' \alpha_0, x) \gamma_0 \dot{h}(t + x' \alpha_0) \{y - \dot{b}(D'(t + x' \alpha_0) \theta_0)\} / a(\phi_0) \right. \\
&\quad \left. \left(\{1 - \eta_0(t)\} \{m_1(\alpha_0, \epsilon_0; t) + m_2(\alpha_0, \epsilon_0; t, \Delta)\} + \dot{F}_\alpha(\alpha, t) m_3(\alpha_0, \epsilon_0; x, \Delta) \right) dt \right\} \\
&\quad \left(\int_{C - x' \alpha_0}^{\tau} f_{\theta_0, \phi_0}(y \mid t + x' \alpha_0, x) \{y - \dot{b}(D'(t + x' \alpha_0) \theta_0)\} D(t + x' \alpha_0) d\eta_0(t) \right) \\
&\quad \left(\int_{C - x' \alpha_0}^{\tau} f_{\theta_0, \phi_0}(y \mid t + x' \alpha_0, x) d\eta_0(t) \right)^{-2} dy dF_X(x) \Big] + o_p(1) \\
&= \mathbb{G}_n \{ G_2(\theta_0, \phi_0, \alpha_0, \eta_0, \Delta, Y, X, V) \}
\end{aligned} \tag{23}$$

Hence, $(\Psi_{n,\theta_0} - \Psi_{\theta_0})(\phi_0, \alpha_0, \eta_0) + \dot{\Psi}_{2,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\phi}_n - \phi_0) + \dot{\Psi}_{3,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\alpha}_n - \alpha_0) + \dot{\Psi}_{4,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\eta}_{n,\hat{\alpha}_n} - \eta_0)$ is the sum of independent and identically distributed terms and the classical central limit theorem applies. We have $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges weakly to a mean zero normal random variable with variance $A^{-1}BA^{-1}$, where

$$\begin{aligned}
A &= -\dot{\Psi}_{1,\theta_0}(\phi_0, \alpha_0, \eta_0), \\
B &= \left\{ G_1(\theta_0, \phi_0, \alpha_0, \eta_0, \Delta, Y, X, V) + \dot{\Psi}_{2,\theta_0}(\phi_0, \alpha_0, \eta_0) m_4(\theta_0, \Delta, Y, X, V) \right. \\
&\quad \left. + \dot{\Psi}_{3,\theta_0}(\phi_0, \alpha_0, \eta_0) m_3(\alpha_0, \epsilon_0; \Delta, X) + G_2(\theta_0, \phi_0, \alpha_0, \eta_0, \Delta, Y, X, V) \right\}^{\otimes 2}.
\end{aligned}$$

Note that for other rank based estimates of α , m_3 in B is the corresponding influence function with different forms; For the sieve maximum likelihood estimates (Ding and Nan, 2011), m_3 is the efficient influence function (Ritov and Wellner, 1988). It is clearly seen that the analytic form of the asymptotic variance is too complicated to be useful for the asymptotic variance estimation. Hence in our numerical studies we use bootstrap to obtain the variance estimator. \square

References

DING, Y., AND NAN, B. (2011). A sieve M-theorem for bundled parameters in

- semiparametric models, with application to the efficient estimation in a linear model for censored data. *Annals of Statistics* **39**, 3032-3061.
- NAN, B. AND WELLNER J. A. (2013). A general semiparametric Z-estimation approach for case-cohort studies. *Statistica Sinica*. **23(3)** 1155-1180.
- RITOV, Y. & WELLNER J. A. (1988). Censoring, martingale, and the Cox model, Contemporary Mathematics. *Statistical Inference for Stochastic processes*, ed. N.H. Prabhu. **80** 191-219.
- VAN DER VAART, A. W. & WELLNER, J.A. (1996). Weak Donvergence and Empirical Process. *Springer*.